

EXHIBIT A



US007072849B1

(12) **United States Patent**
Filepp et al.

(10) **Patent No.:** US 7,072,849 B1
(45) **Date of Patent:** Jul. 4, 2006

(54) **METHOD FOR PRESENTING ADVERTISING IN AN INTERACTIVE SERVICE**

(75) Inventors: **Robert Filepp**, White Plains, NY (US); **Alexander W. Bidwell**, New York, NY (US); **Francis C. Young**, Pearl River, NY (US); **Allan M. Wolf**, Ridgefield, CT (US); **Duane Tiemann**, Ossining, NY (US); **Mel Bellar**, New York, NY (US); **Robert D. Cohen**, Pouyhquag, NY (US); **James A. Galambos**, deceased, late of Westport, CT (US); **Kenneth H. Appleman**, Brewster, NY (US); **Sam Meo**, Carmel, NY (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: 08/158,025

(22) Filed: Nov. 26, 1993

Related U.S. Application Data

(60) Division of application No. 07/388,156, filed on Jul. 28, 1989, now Pat. No. 5,347,632, which is a continuation-in-part of application No. 07/328,790, filed on Mar. 23, 1989, now abandoned, which is a continuation-in-part of application No. 07/219,931, filed on Jul. 15, 1988, now abandoned.

(51) **Int. Cl.**
G06Q 30/00 (2006.01)

(52) **U.S. Cl.** 705/14
(58) **Field of Classification Search** 364/401; 395/600, 144, 153, 200, 250, 201, 207, 210, 395/214, 611, 613, 614, 615, 762, 779, 782, 395/133, 135, 507, 327, 339, 340, 343, 346, 395/200.09, 445, 460

See application file for complete search history.

(56)

References Cited

U.S. PATENT DOCUMENTS

3,653,001 A * 3/1972 Ninke 395/132

(Continued)

FOREIGN PATENT DOCUMENTS

JP 573167 * 1/1982
JP 3204259 * 9/1991

OTHER PUBLICATIONS

"Trintex Sets Prodigy Pricing; Telaction Reports New Cable System Affiliate"; *IDP Report*; v 9 Issue:n4 p. 2(2); Apr. 1, 1988; Dialog(file 648, 06639981).*

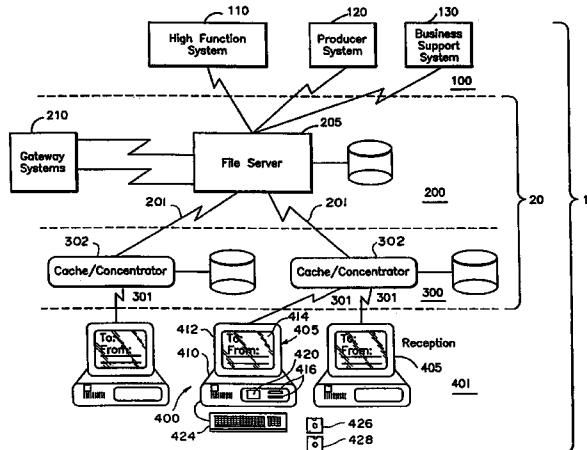
(Continued)

Primary Examiner—Donald L. Champagne*(74) Attorney, Agent, or Firm*—Connolly Bove Lodge & Hutz LLP; Douglas Lefevre

(57)

ABSTRACT

A method for presenting advertising in an interactive service provided on a computer network, the service featuring applications which include pre-created, interactive text/graphic sessions is described. The method features steps for presenting advertising concurrently with service applications at the user terminal configured as a reception system. In accordance with the method, the advertising is structured in a manner comparable to the service applications enabling the applications to be presented at a first portion of a display associated with the reception system and the advertising presented at a second portion. Further, steps are provided for storing and managing advertising at the user reception system so that advertising can be pre-fetched from the network and staged in anticipation of being called for presentation. This minimizes the potential for communication line interference between application and advertising traffic and makes the advertising available at the reception system so as not to delay presentation of the service applications. Yet further the method features steps for individualizing the advertising supplied to enhance potential user interest by providing advertising based on a characterization of the user as defined by the users interaction with the service, user demographics and geographical location. Yet additionally, advertising is provided with transactional facilities so that users can interact with it.

25 Claims, 16 Drawing Sheets

US 7,072,849 B1

Page 2

U.S. PATENT DOCUMENTS

4,552,349 A *	11/1985	Loos et al.	270/54
4,575,579 A *	3/1986	Simon et al.	178/4
4,688,167 A *	8/1987	Agarwal	395/343
4,714,996 A *	12/1987	Gladney et al.	395/600
4,805,134 A *	2/1989	Calo et al.	395/610
4,823,122 A *	4/1989	Mann et al.	340/825.28
4,873,662 A	8/1989	Sargent	
4,887,204 A *	12/1989	Johnson et al.	395/600
4,897,781 A *	1/1990	Chang et al.	395/600
4,897,782 A *	1/1990	Bennett et al.	395/600
4,989,850 A *	2/1991	Weller	270/1.1
5,036,314 A *	7/1991	Barillari et al.	340/717
5,087,805 A *	2/1992	Silverschotz et al. ...	219/121.71
5,105,184 A *	4/1992	Pirani et al.	340/721
5,119,290 A *	6/1992	Loo et al.	395/400

OTHER PUBLICATIONS

- "The Handbook"; Prodigy; ©1990 Prodigy Services Company Glessbrenner, Alfred; Ceriés, New On-line fee; \$4.95 a Month; *Home Office Computing*; v8 P. 36(1); Dec., 1990 Dialog (file 647, 09685321).*
- "Advertisers Need Quick Fix for Zipping, Zapping"; *Marketing News*; v20 n10; pp. 12; May 9, 1986; Dialog: File 15, Acc# 00317906.*
- "Consumers Plugging into new Electronic Mall"; *Advertising Age*; Mar. 4, 1985; p. 741; Dialog: File 16, Acc# 01155574.*
- "CompuServe Will Jointly Offer Advertising and Direct Marketing Services via the CompuServe Information Service, a Videotex System"; *News Release*; Oct. 19, 1983; pp. 1-3; Dialog: File 16, Acc# 00962377.*
- "Compuserve, L.M. Berry to Test Viability of Online Advertising"; *Online Database Report*; v4 n10; p. 12; Oct. 1983; Dialog: File 275, Acc# 00610155.*
- Miller; "Database and Videotex Services—Where is Videotex Going?"; *Data Communications Buyers' Guide 1983*; pp. 152-157; Nov. 1982; Dialog: File 15, Acc# 00188062.*
- Dictionary of Computers, Information Processing & Telecommunications, 2nd ed.*; Jerry M Rosenberg; 1984; pp. 700.*
- Miller; "Database and Videotex Services—Where Is Videotex Going?"; *Data Communications Buyers' Guide 1983*; pp. 152, 157; Nov. 1982.*
- Dietrich et al.; "Toward a Graphic Standard"; *PC World*; v2 n12; p. 264-269; Nov. 1984.*
- "MCTel Inc. Advertises in the Electronic Mall Shop-at-Home Service, an Advertising Vehicle of CompuServe Inc. And L. M. Berry & Co."; *PR Newswire, PH303*; Jan. 23, 1985; Dialog: File 148, Acc# 02341095.*
- "Consumers Plugging Into New Electronic Mall"; *Advertising Age*; Mar. 4, 1985; p. 741.*
- "Home-Computer Shopping Arrives"; *Discount Store News*; v24; p. 3(2); Mar. 18, 1985; Dialog: File 148, Acc# 02324097.*
- "Advertisers Need Quick Fix for Zipping, Zapping"; *Marketing News*; v20 n10; pp. 12; May 9, 1986.*
- Caplinger, Michael, "An Information System Based on Distributed Objections", OOPSLA '87 Proceedings.
- Schatz, Bruce, "Telesophy: A System for Manipulating the Knowledge of a Community", 1987 IEEE.
- Christodoulakis, S., "The Multimedia Object Presentation Manager of MINOS: A Symmetric Approach", ACM SIGMOD Conf. 1986.
- Christodoulakis, S., "Issues in the Architecture of a Document Archiver Using Optical Disk Technology", 1985 ACM.
- Christodoulakis, S., "Multimedia Document Presentation, Information Extraction, and Document Formation in MINOS: A Model and A System" 1986 ACM.
- Sigel, Efrem, "The Future of Videotext", 1983, Knowledge Industry Publications, Inc., White Plains NY and London.
- Alber, Antone F., *Videotex/Teletext Principles & Practices*, McGraw-Hill, Inc., 1985.

* cited by examiner

U.S. Patent

Jul. 4, 2006

Sheet 1 of 16

US 7,072,849 B1

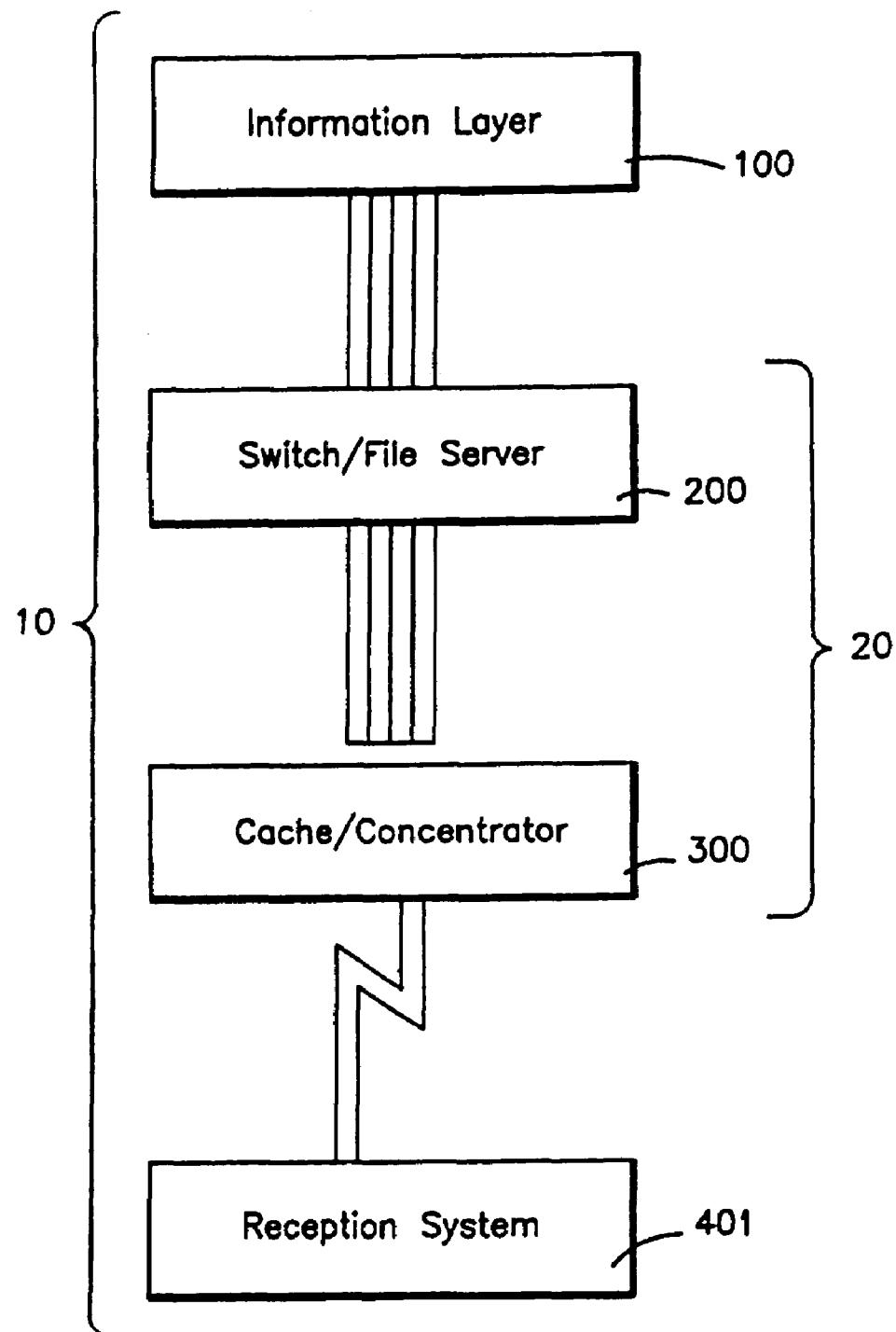


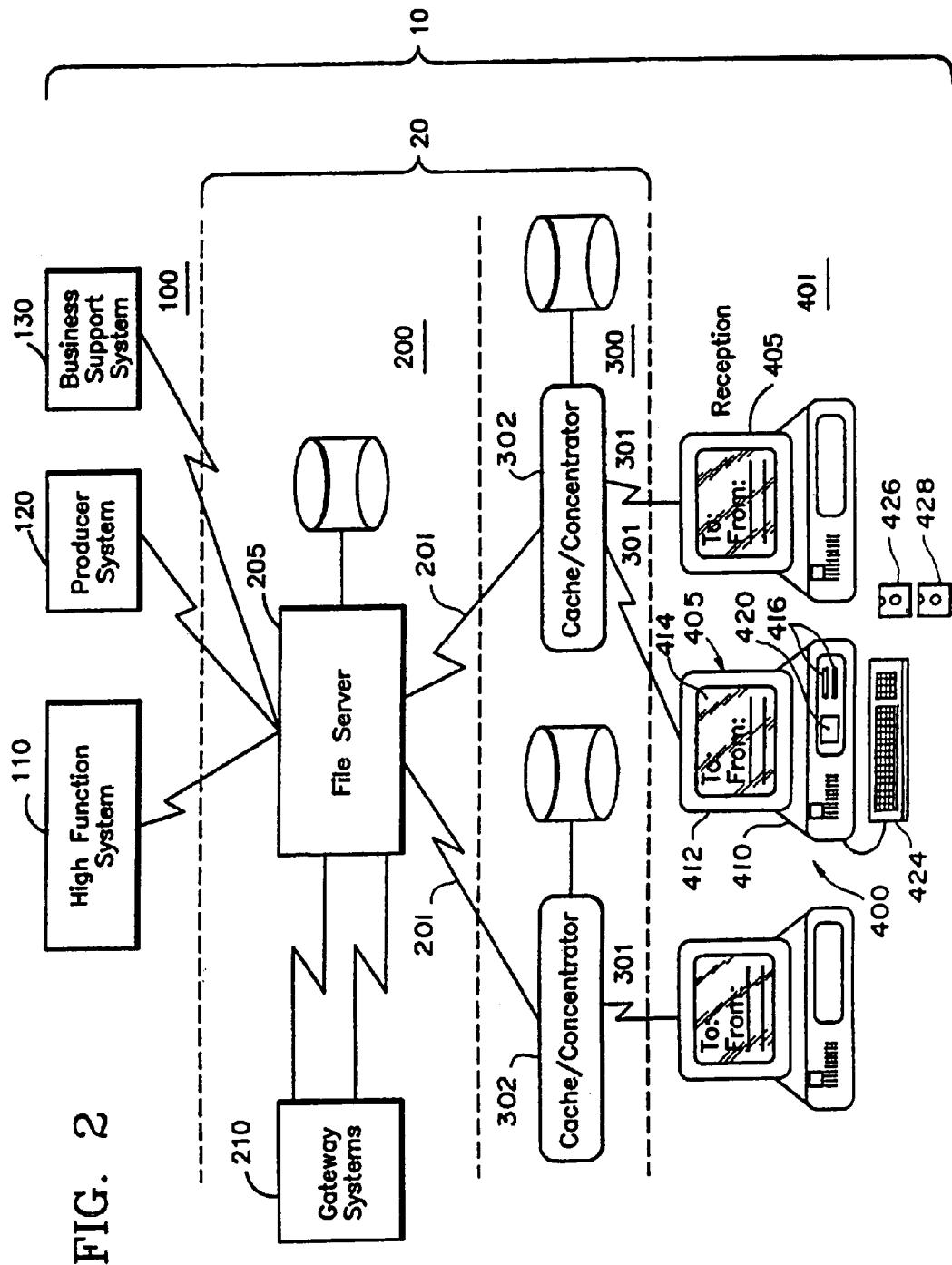
FIG. 1

U.S. Patent

Jul. 4, 2006

Sheet 2 of 16

US 7,072,849 B1



U.S. Patent

Jul. 4, 2006

Sheet 3 of 16

US 7,072,849 B1

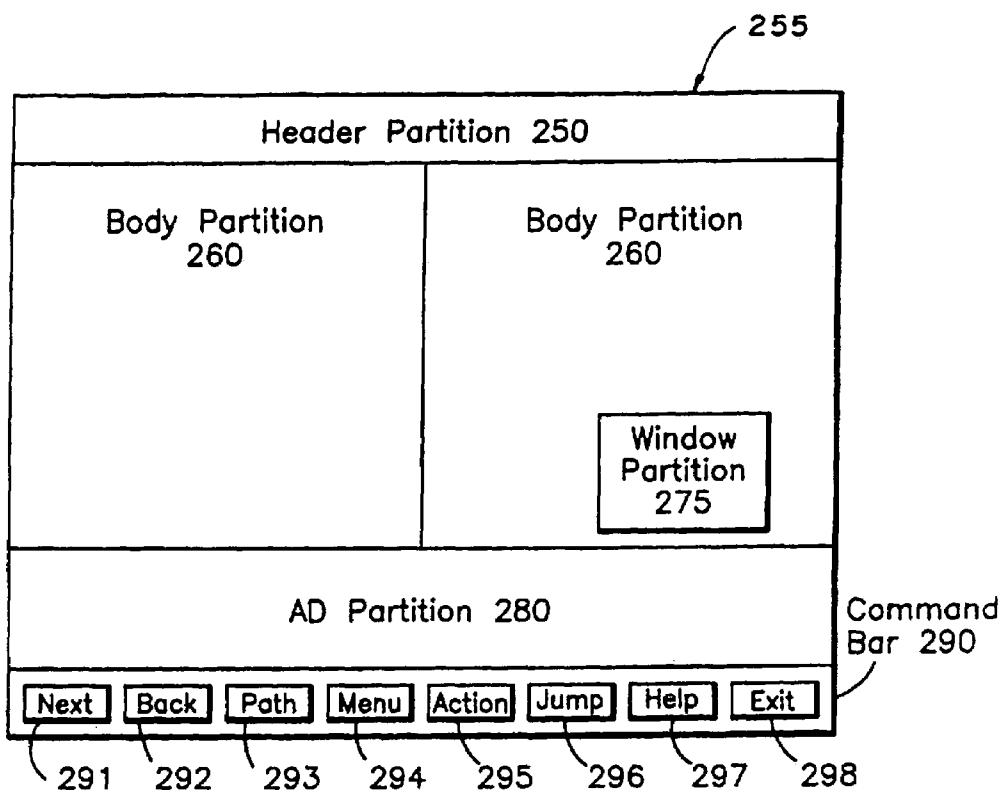


FIG. 3a

U.S. Patent

Jul. 4, 2006

Sheet 4 of 16

US 7,072,849 B1

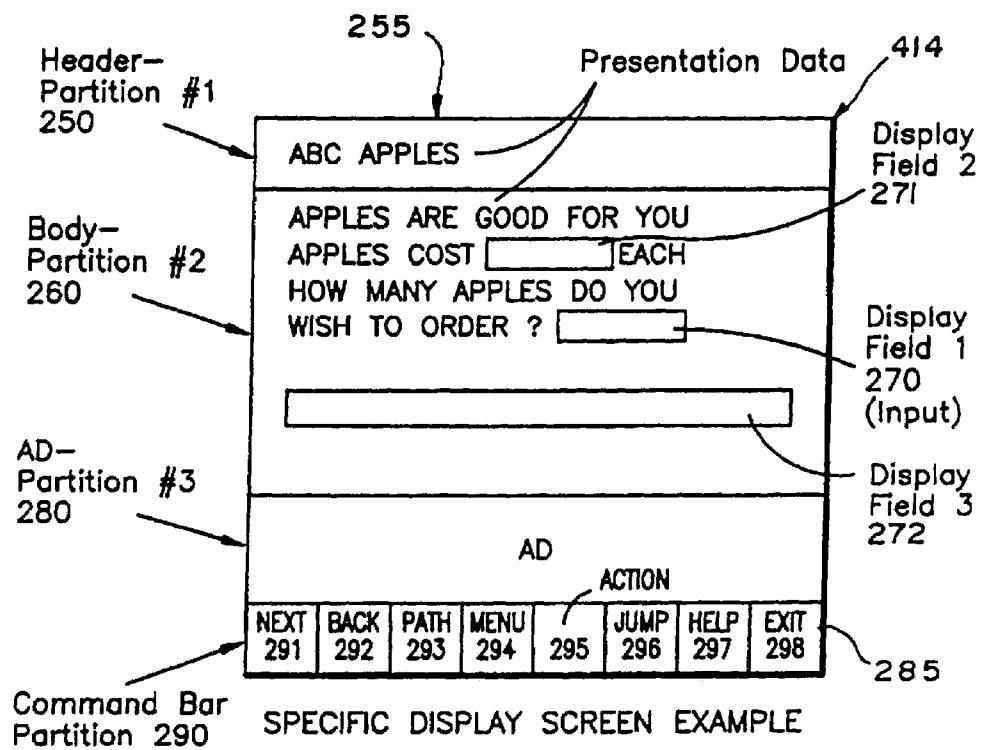


FIG. 3b

U.S. Patent

Jul. 4, 2006

Sheet 5 of 16

US 7,072,849 B1

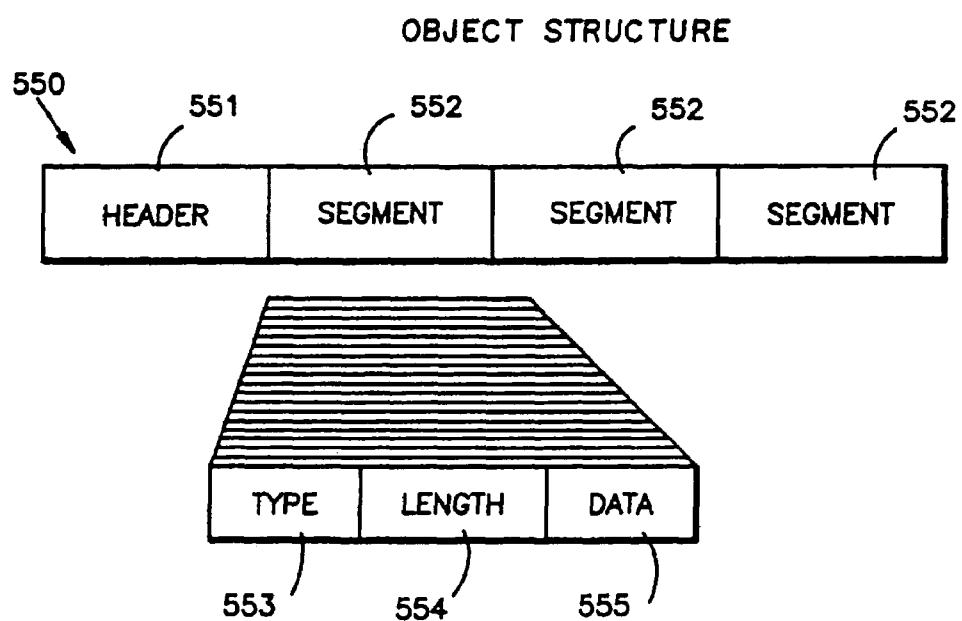


FIG. 4a

U.S. Patent

Jul. 4, 2006

Sheet 6 of 16

US 7,072,849 B1

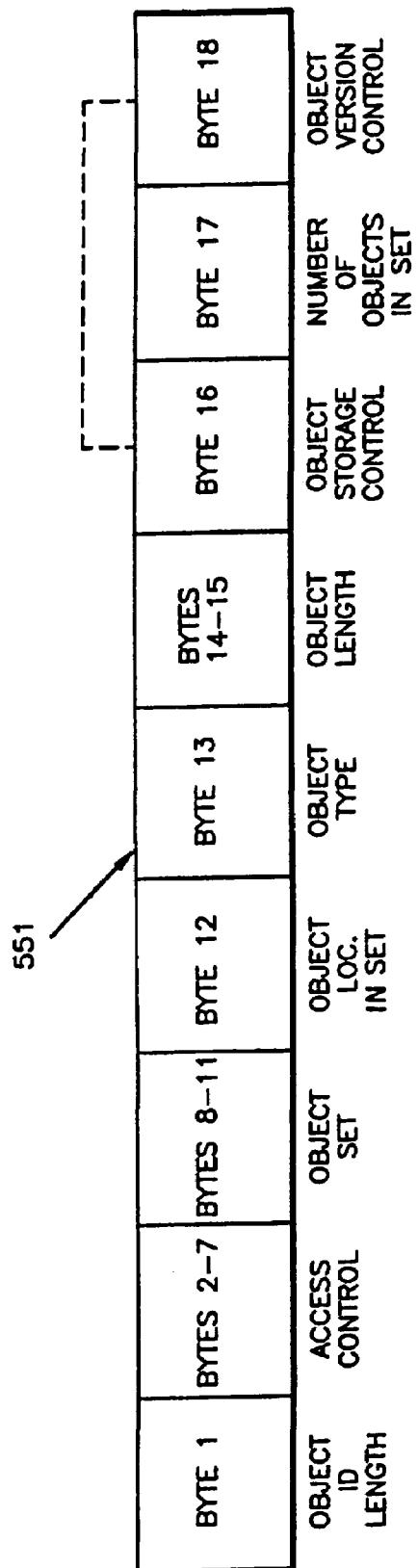


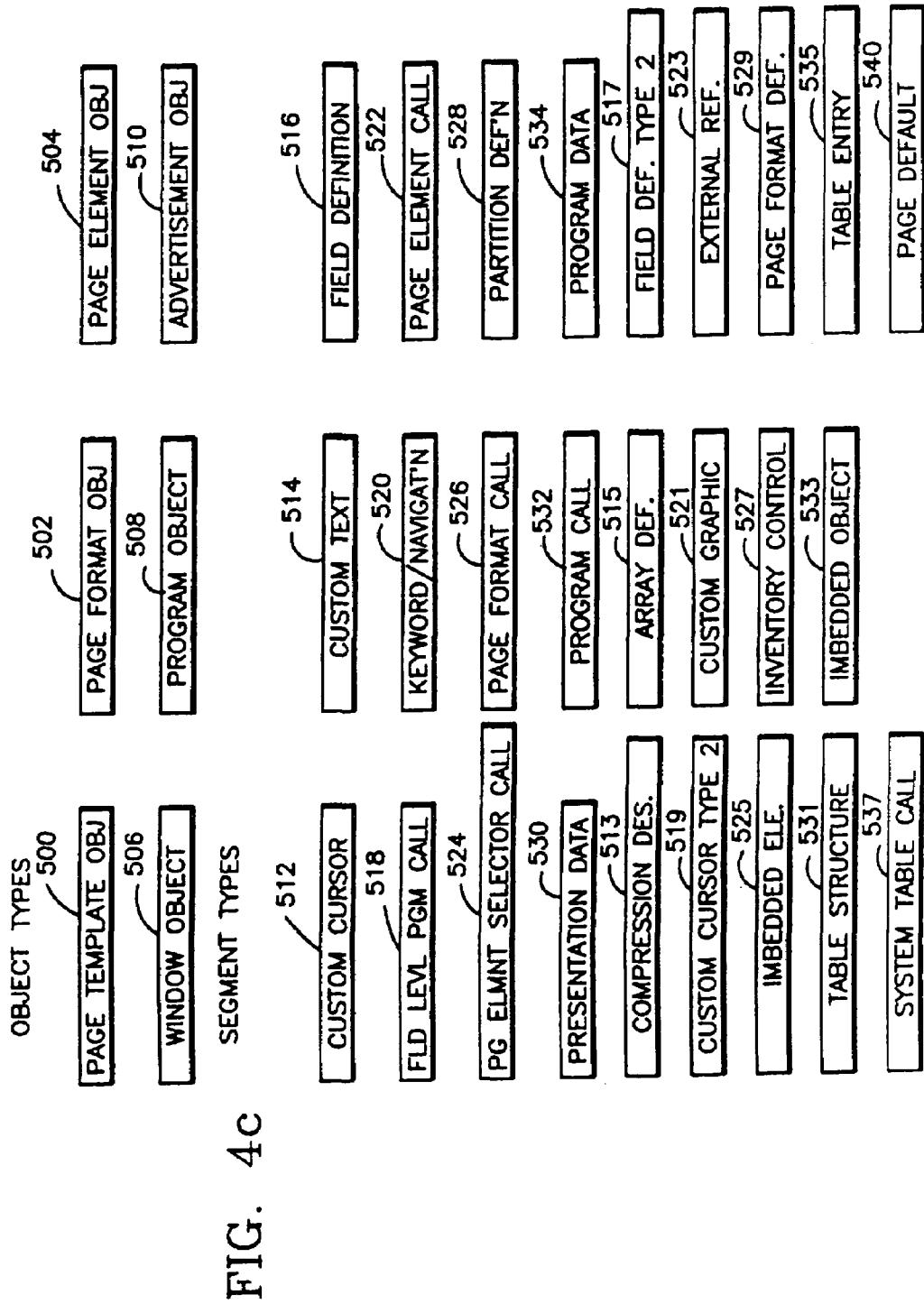
FIG. 4b

U.S. Patent

Jul. 4, 2006

Sheet 7 of 16

US 7,072,849 B1



U.S. Patent

Jul. 4, 2006

Sheet 8 of 16

US 7,072,849 B1

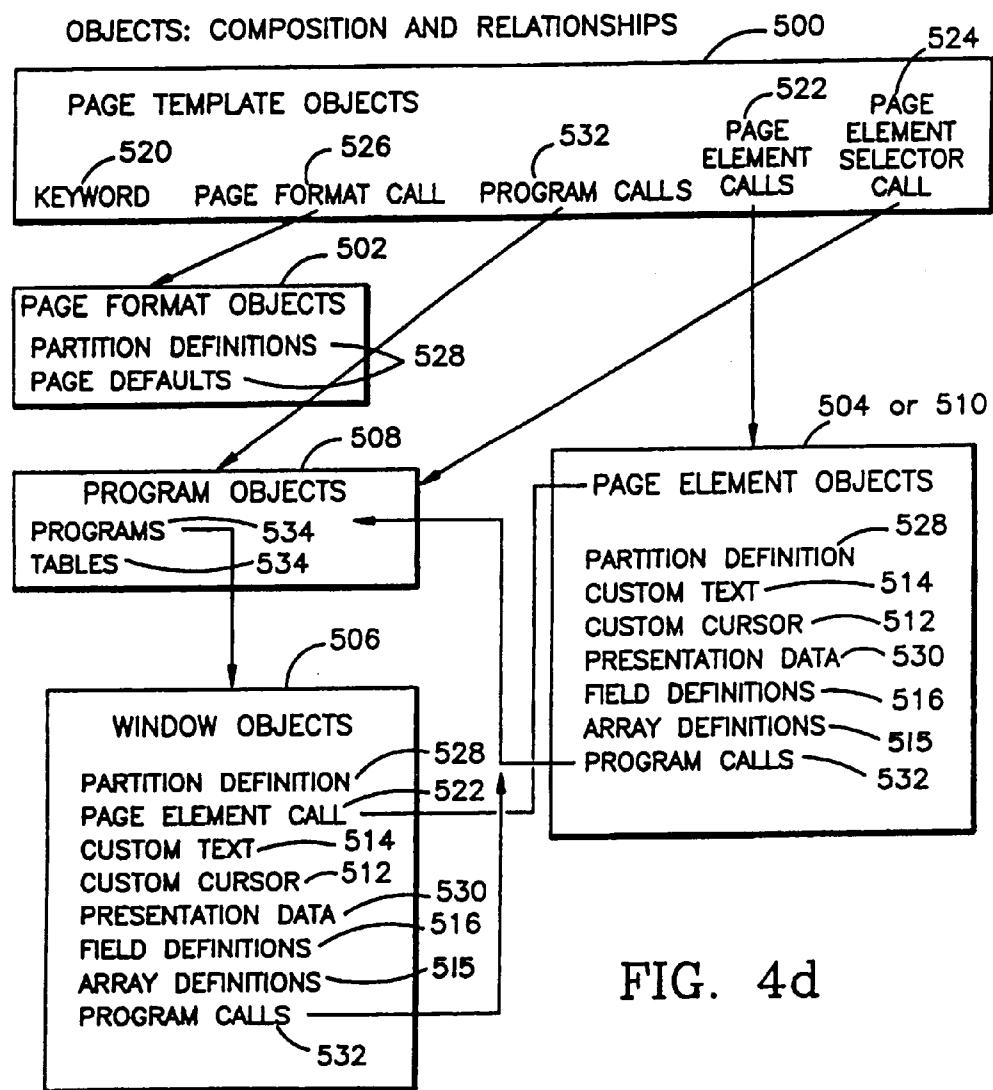


FIG. 4d

U.S. Patent

Jul. 4, 2006

Sheet 9 of 16

US 7,072,849 B1

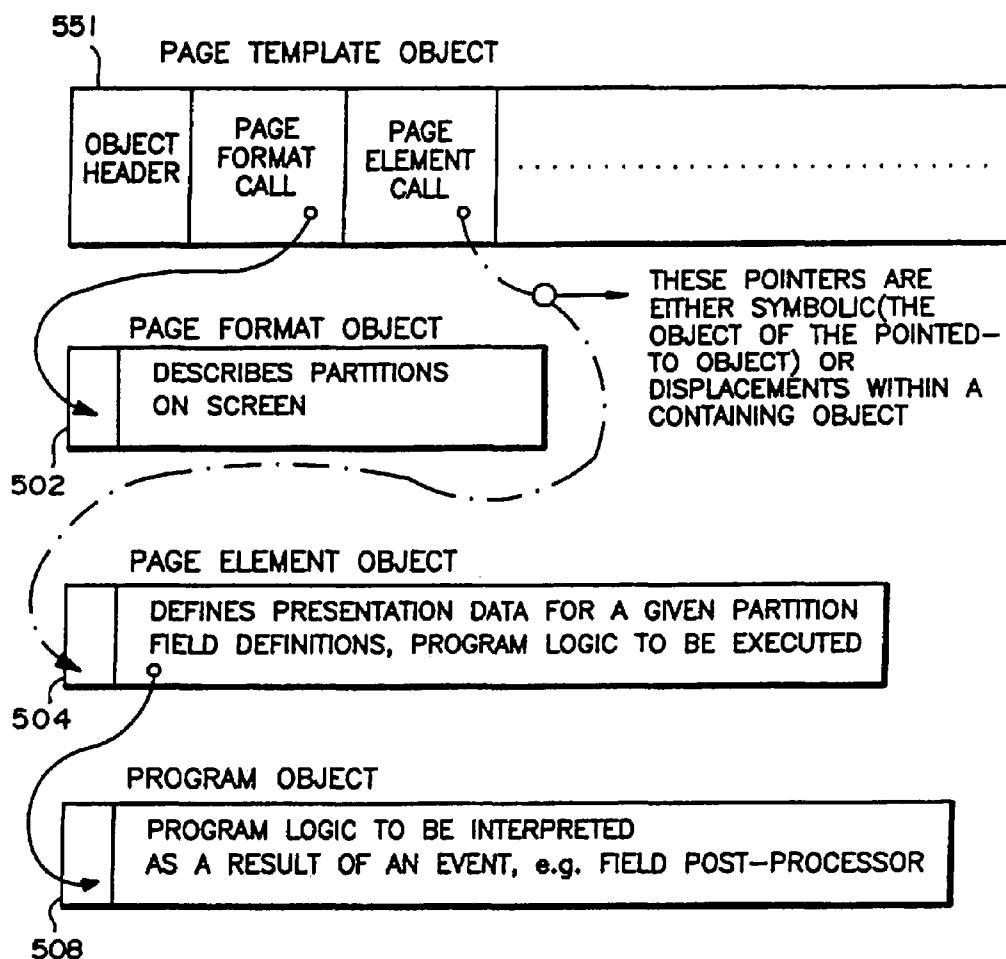


FIG. 5a

U.S. Patent

Jul. 4, 2006

Sheet 10 of 16

US 7,072,849 B1

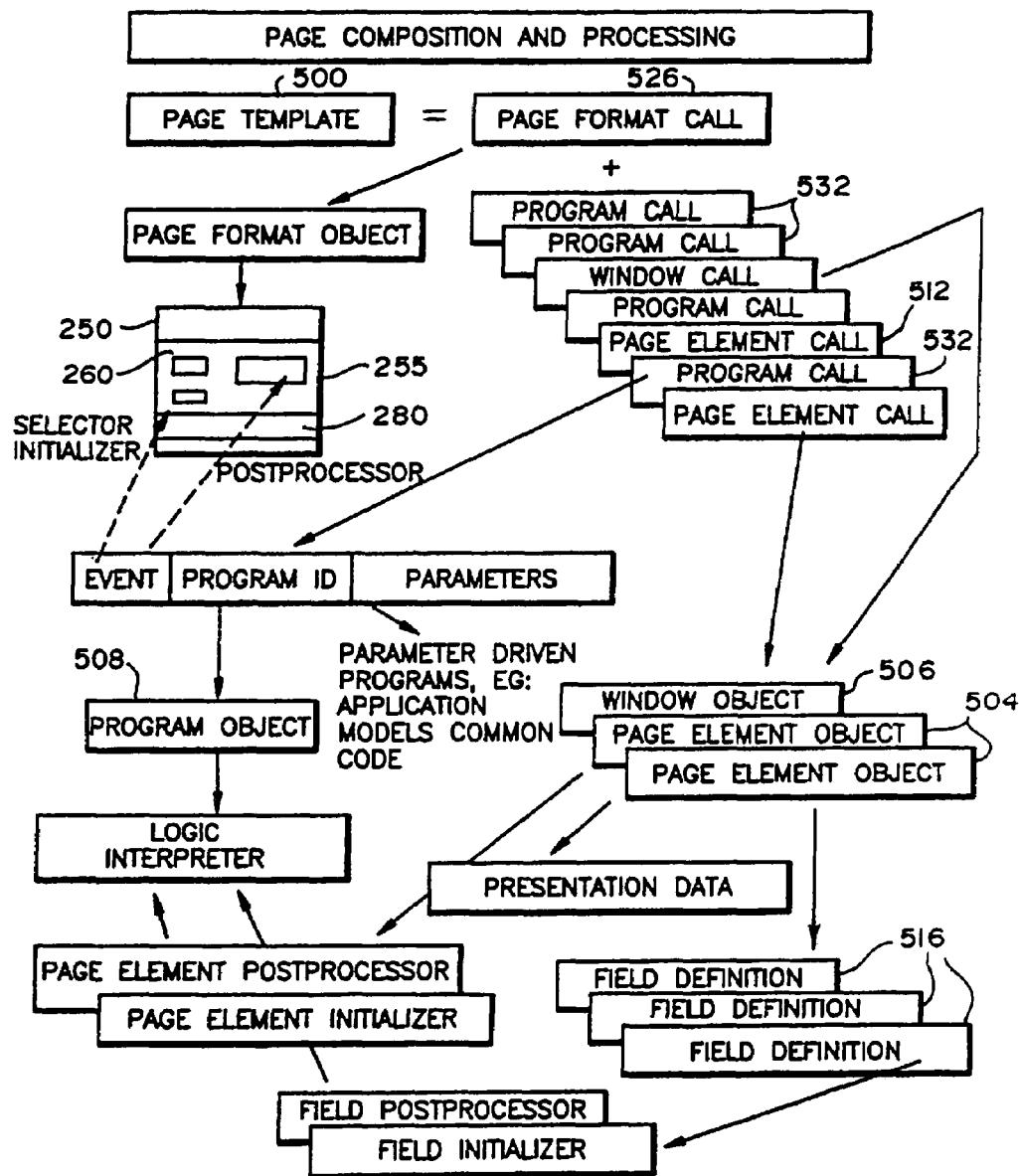


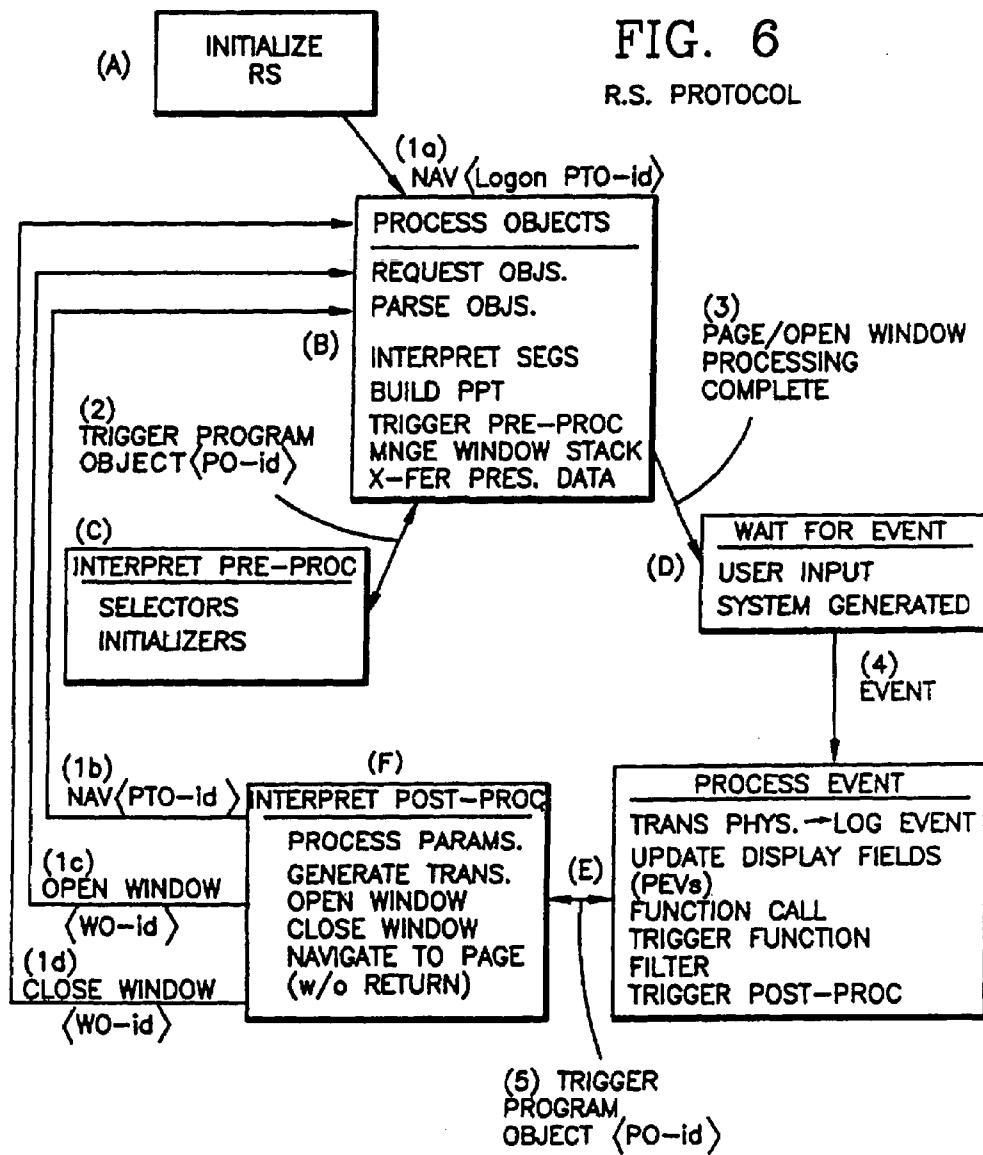
FIG. 5b

U.S. Patent

Jul. 4, 2006

Sheet 11 of 16

US 7,072,849 B1



U.S. Patent

Jul. 4, 2006

Sheet 12 of 16

US 7,072,849 B1

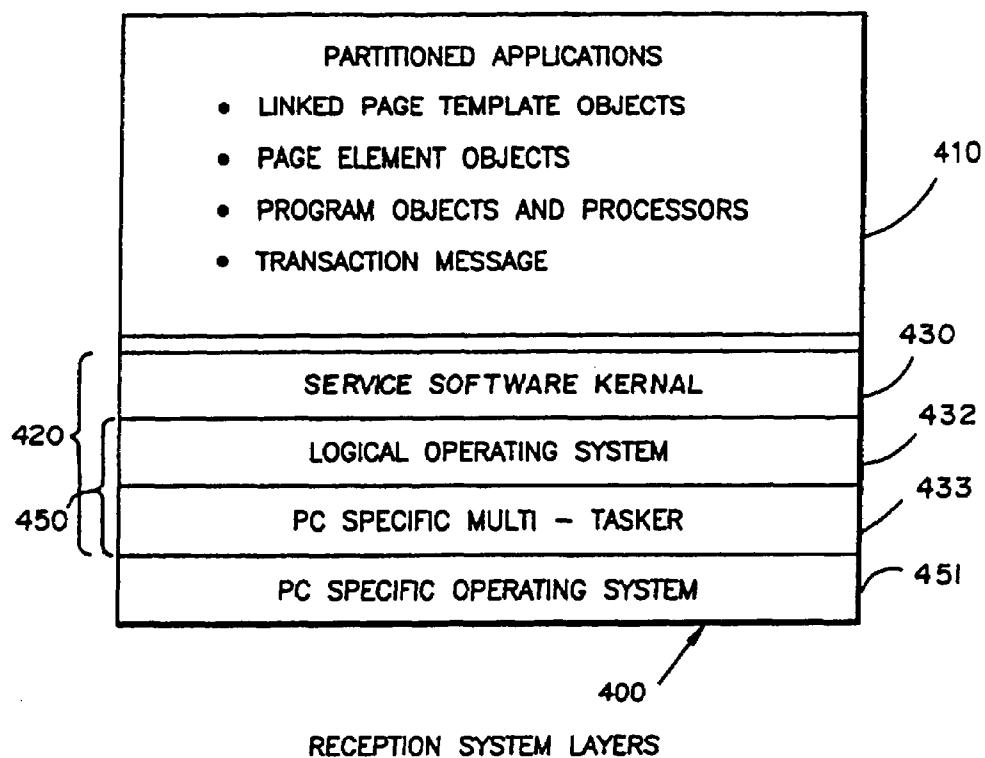


FIG. 7

U.S. Patent

Jul. 4, 2006

Sheet 13 of 16

US 7,072,849 B1

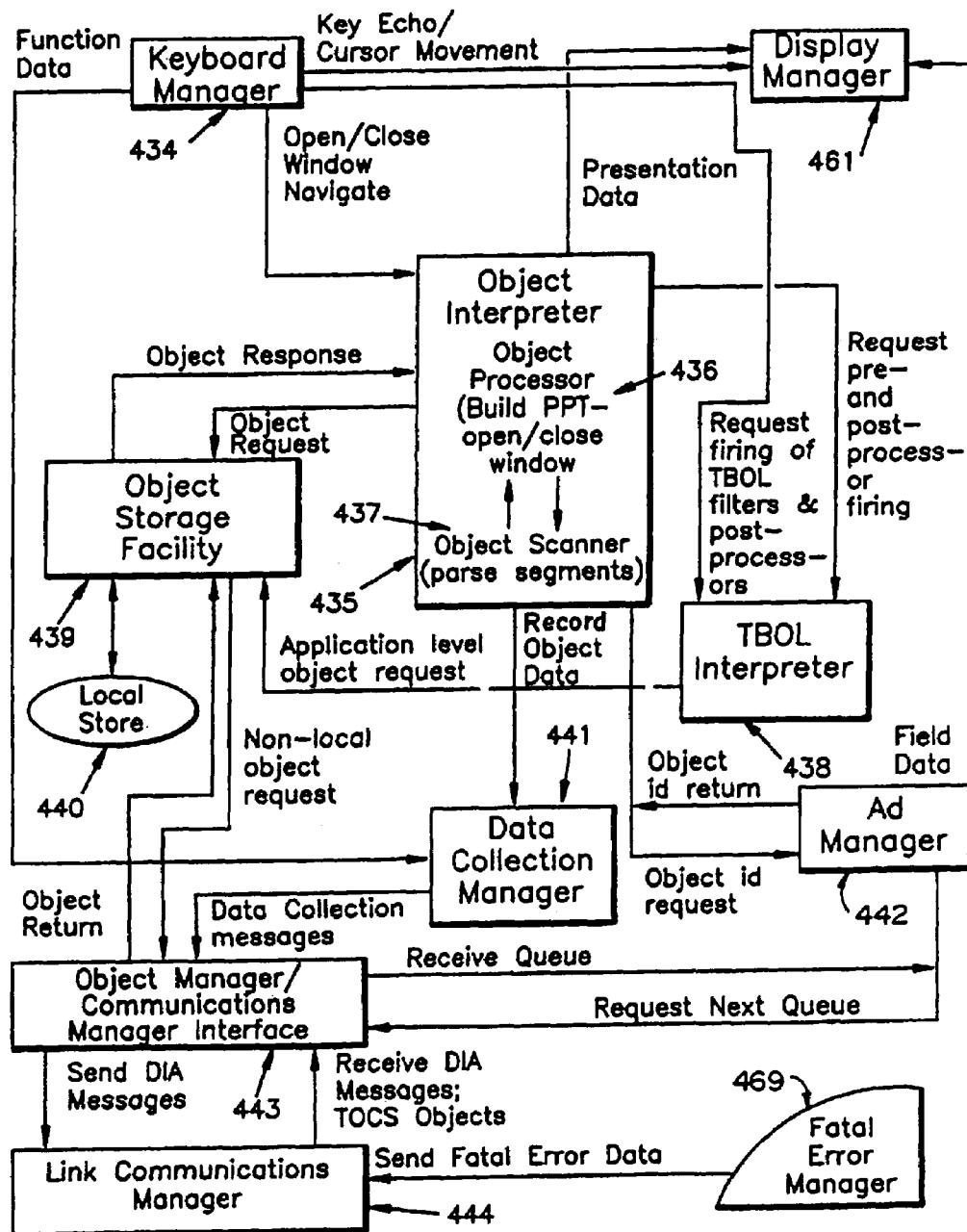


FIG. 8

U.S. Patent

Jul. 4, 2006

Sheet 14 of 16

US 7,072,849 B1

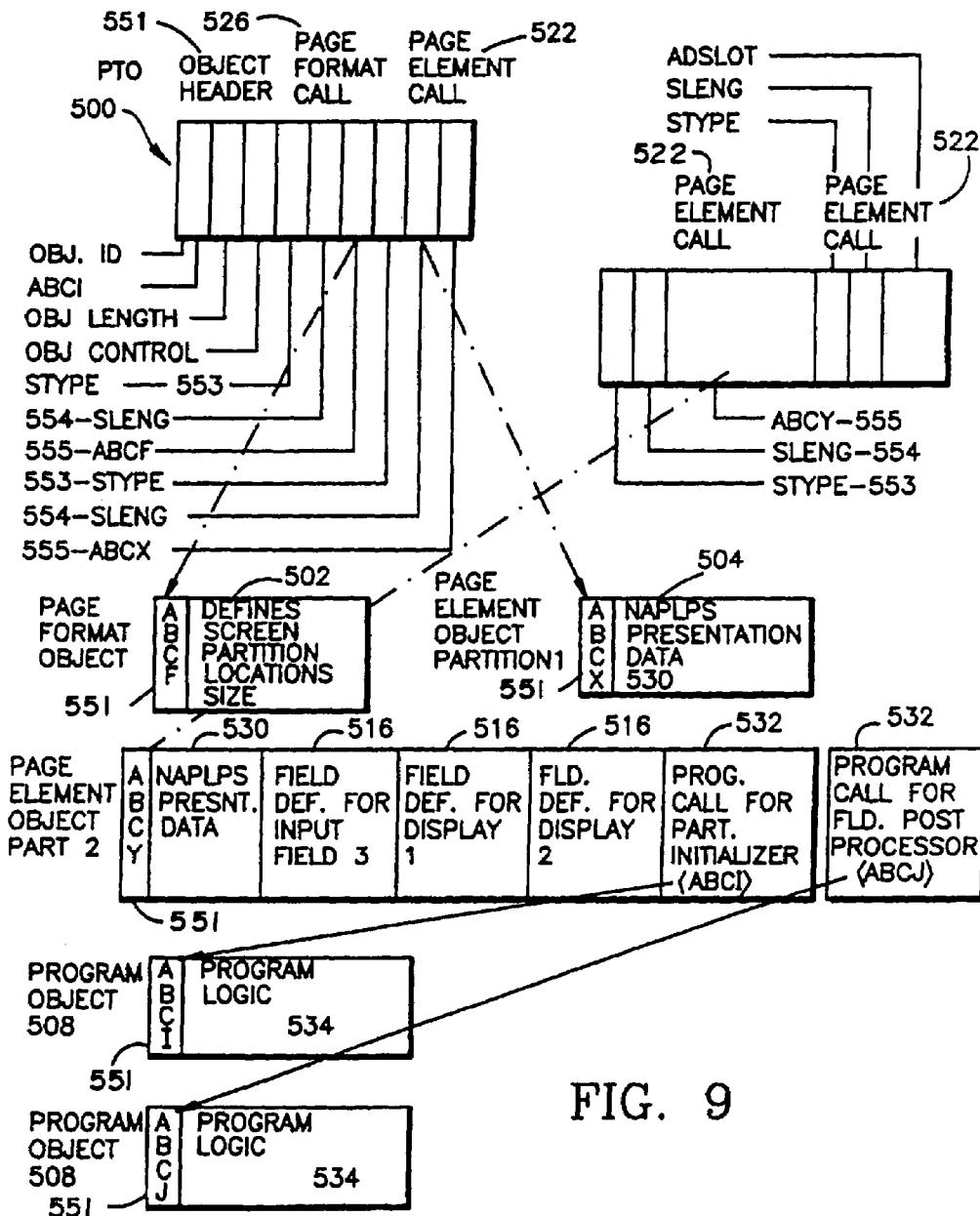


FIG. 9

U.S. Patent

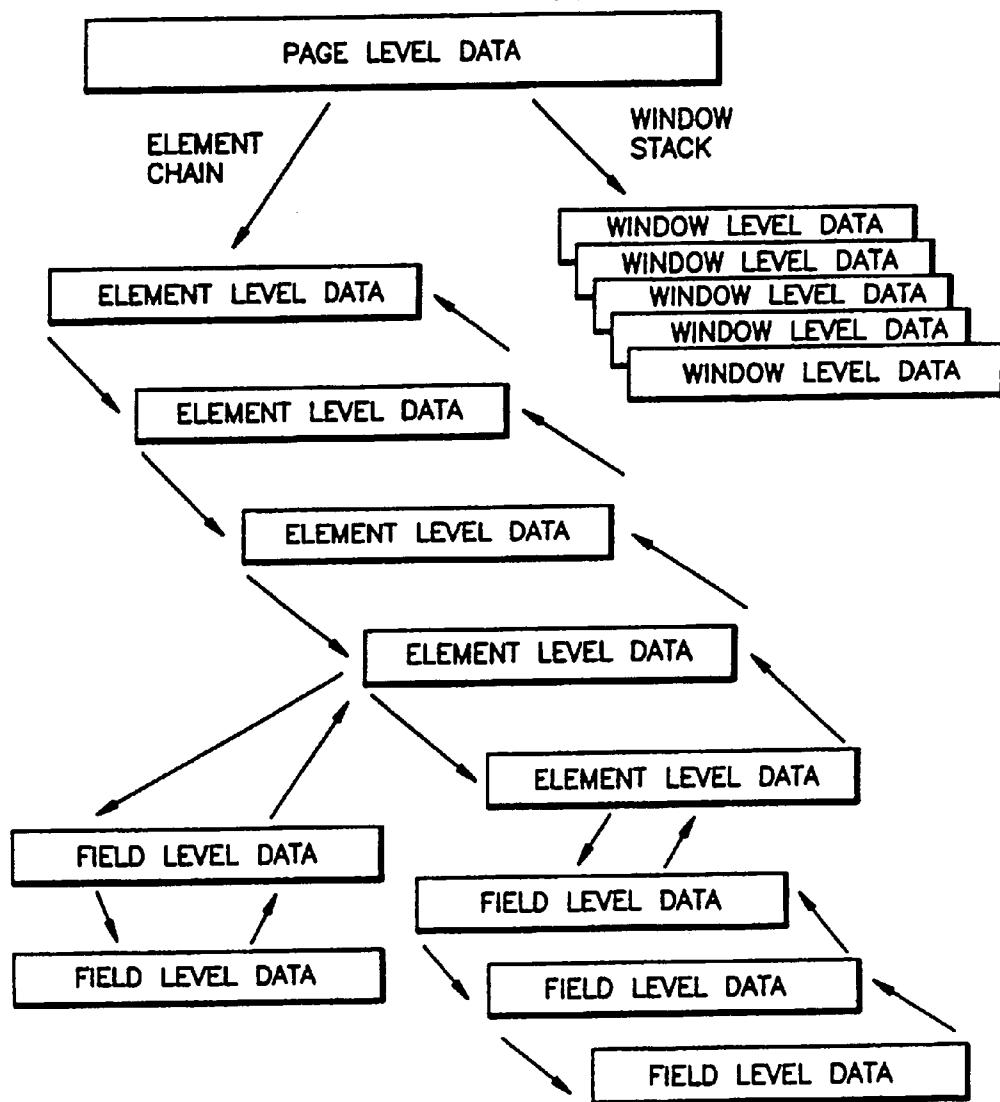
Jul. 4, 2006

Sheet 15 of 16

US 7,072,849 B1

FIG. 10

PAGE PROCESSING TABLE (ppt)



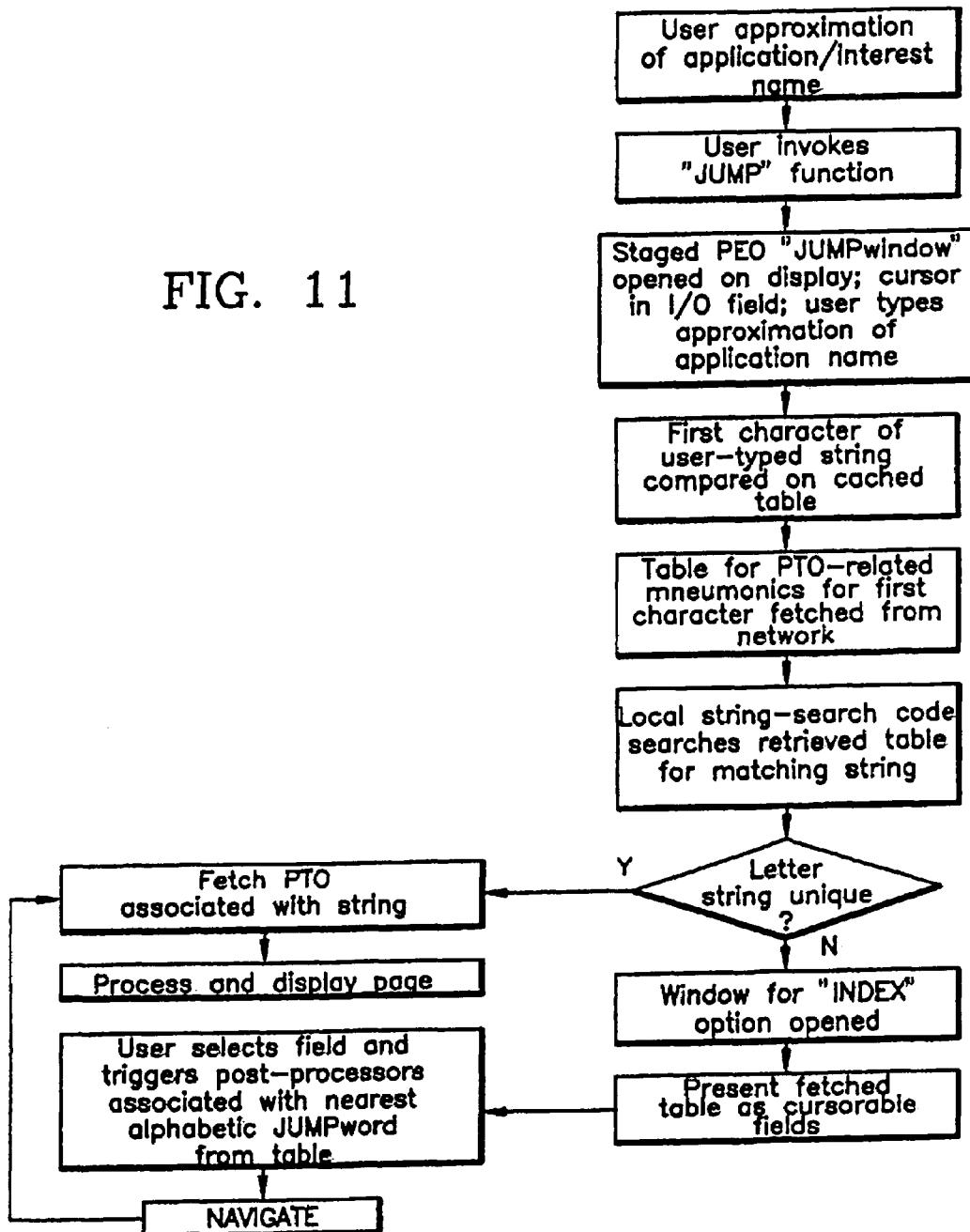
U.S. Patent

Jul. 4, 2006

Sheet 16 of 16

US 7,072,849 B1

FIG. 11



US 7,072,849 B1

1**METHOD FOR PRESENTING ADVERTISING
IN AN INTERACTIVE SERVICE****RELATED APPLICATIONS**

This is a division of application Ser. No. 07/388,156 filed Jul. 28, 1989, Sep. 13, 1994, as U.S. Pat. No. 5,347,632, application Ser. No. 07/388,156 being a continuation in part of application Ser. No. 07/328,790, now abandoned filed Mar. 23, 1989, which itself was a continuation in part of application Ser. No. 07/219,931, now abandoned filed Jul. 15, 1988.

BACKGROUND OF THE INVENTION**1. Field of Use**

This invention relates generally to a distributed processing, interactive computer network intended to provide very large numbers of simultaneous users; e.g. millions, access to an interactive service having large numbers; e.g., thousands, of applications which include pre-created, interactive text/graphic sessions; and more particularly, to a method for presenting advertising to service users during interactive sessions, the method featuring steps for presenting advertising concurrently with applications, the advertising being organized as data which is stored for presentation and replenished at the user sites so as to minimize interference with retrieval and presentation of application data; the method also featuring steps for individualizing the advertising presented based on user characterizations defined by service interaction and/or other data such as user demographics and geographical location.

2. Prior Art

Interactive computer networks are not new. Traditionally they have included conventional, hierarchical architectures wherein a central, host computer responds to the information requests of multiple users. An illustration would be a time-sharing network in which multiple users, each at a remote terminal, log onto a host that provides data and software resource for sequentially receiving user data processing requests, executing them and supplying responses back to the users.

While such networks have been successful in making the processing power of large computers available to many users, problems have existed with them. For example, in such networks, the host has been required to satisfy all the user data processing requests. As a result, processing bottlenecks arise at the host that cause network slowdowns and compel expansion in computing resources; i.e., bigger and more complex computer facilities, where response times are sought to be held low in the face of increasing user populations.

Host size and complexity, however, are liabilities for interactive networks recently introduced to offer large numbers of the public access to transactional services such as home shopping, banking, and investment maintenance, as well as informational services concerning entertainment, business and personal matters. As can be appreciated, commercial interactive networks will have to provide attractive services at low cost and with minimal response times in order to be successful. Unlike military and governmental networks where, because of the compulsory nature of the service performed, costs, content and efficiency are of secondary concern, in commercial services, since use is predominantly elective, and paid for by the consumer, costs will have to be held low, content made interesting and response times reduced in order to attract and hold both users who

2

would subscribe to the service and merchandisers who would rely on it as a channel of distribution for their good and services. Accordingly, if the service delivery system is allowed to increase in size and complexity, either unchecked or unsubsidized, higher use costs would have to be charged to recover the larger capital and operating expenses, with the negatively, spiralling effect that fewer users could be attracted and be available over which to spread the costs for sustaining the service.

In the past, other suppliers of mass-media services such as radio, television, newspapers, and magazines, have sought to hold access and subscription prices to affordable levels by relying on advertising income to offset the costs of providing their users with the benefits of technological advance. However, in the case of interactive computer services, it has not been apparent how advertising could be introduced without adversely affecting service speed and content quality, which as noted, are considered essential elements for service success.

Particularly, in an interactive service, if advertising were provided in a conventional manner; as for example, by providing the advertising as additional data to be supplied to and presented at the user sites, the effort would compete with the supplying and presentation of service application data, and have the undesirable effect of diminishing service response time. More specifically, if advertising were supplied conventionally from a host to a user site, the application traffic, which constitutes the substance of the service, would have to compete with advertising for network communication resources. Yet additionally, even if traffic conflicts were somehow avoided, the presentation of the service applications would have to be interrupted and delayed; for example like television and radio commercials, as advertising content was presented to the user. The effect of these anticipated delays would be to degrade application response time and diminishing service attractiveness.

Additionally, in view of the need to maintain the user's interest in application content so as to drive the interactive session, it has not been apparent how advertising matter could be provided without distracting the user or disrupting the session. Where service response time is diminished for the sake of advertising which is either irrelevant or distasteful, insult is added to the injury, increasing the likelihood the user, and service, will be, simply, turned off.

SUMMARY OF INVENTION

Accordingly, it is an object of this invention to provide a method for presenting advertising in an interactive service.

It is another object of this invention to provide a method for presenting advertising in an interactive service which method enables the presentation of advertising to be integrated with presentation of service applications.

It is a yet another object of this invention to provide a method for presenting advertising in an interactive service which method minimizes the potential for interference between the supply of interactive-service applications and advertising.

It is a still another object of this invention to provide a method for presenting advertising which minimizes the potential for interference between presentation of interactive-service applications and advertising. It is yet a further object of this invention to provide a method for presenting advertising in an interactive service which method enables the advertising presented to be individualized to the user to whom it is presented in order to increase the likelihood the advertising will be of interest to the user.

US 7,072,849 B1

3

And, it is still a further object of this invention to provide a method for presenting advertising in an interactive service which method enables the user to transactionally interact with the advertising presented.

Briefly, the method for presenting advertising in accordance with this invention achieves the above-noted and other objects by featuring steps for presenting advertising concurrently with service applications at the user reception system; i.e., terminal. In accordance with the method, the advertising is structured in a manner comparable to the manner in which the service applications are structured. This enables the applications to be presented at a first portion of a display associated with the reception system and the advertising to be presented concurrently at a second portion of the display. Further, in accordance with the method, the user reception system at which the advertising is presented includes facility for storing and managing the advertising so that it can be pre-fetched from the network and staged at the reception system in anticipation of being called for presentation. This minimizes the potential for communication line interference between application and advertising traffic and makes the advertising available at the reception system so as not to delay presentation of the service applications. Yet further the method features steps for individualizing the advertising supplied to enhance potential user interest by providing advertising based on a characterization of the user as defined by the users interaction with the service, user demographics and geographical location. Yet additionally, advertising is provided with transactional facilities so that users can interact with it.

In preferred form, the method includes step for organizing advertising and applications as objects that collectively include presentation data and executable program instructions for generating the advertising and applications at the reception system. In accordance with the preferred form of the method, advertising and application objects are selectively distributed in the service network in accordance with a predetermined plan based on the likelihood the applications and advertising will be called by the respective user reception systems.

Also in preferred form, the method includes step for maintaining an advertising object identification queue, and an advertising object store that are replenished based on predetermined criteria as advertising is called for association and presentation with applications. In accordance with the method, as applications are executed at the reception system, the application objects provide generalized calls for advertising. The application calls for advertising are subsequently forwarded to the reception system advertising queue management facility which, in turn supplies an identification of advertising who's selection has been individualized to the user based on, as noted, the user's prior interaction history with the service, demographics and local. Thereafter, the object identification for the advertising is passed to the object store to determine if the object is available at the reception system. In preferred form, if the advertising object is not available at the reception system, a sequence of alternative advertising object identifications can be provided which if also are unavailable at the reception system will result in an advertising object being requested from the network. In this way, advertising of interest can be targeted to the user and secured in time-efficient manner to increase the likelihood of user interest and avoid service distraction.

4

BRIEF DESCRIPTION OF THE DRAWINGS

The above and further objects, features and advantages of the invention will become clear from the following more detailed description when read with reference to the accompanying drawings in which:

FIG. 1 is a block diagram of the interactive computer network in which the method of the present invention may be practiced;

FIG. 2 is a schematic diagram of the network illustrated in FIG. 1;

FIGS. 3a and 3b are plan views of a display screen for a user reception system at which advertising can be presented to a user in accordance with the method of the present invention;

FIGS. 4a, 4b, 4c and 4d are schematic drawings that illustrate the structure of objects, and object segments that may be used for advertising and applications in accordance with the method of the present invention;

FIG. 5a is a schematic diagram that illustrates the configuration of the page template object which might be used for presentation of an application and advertising in accordance with the method of the present invention;

FIG. 5b is a schematic diagram that illustrates page composition which might be used for presentation of an application and advertising in accordance with the method of the present invention;

FIG. 6 is a schematic diagram that illustrates the protocol which might be used by a reception system for supporting applications and advertising in accordance with the method of the present invention;

FIG. 7 is a schematic diagram that illustrates major layers for a reception system which might be used for supporting applications and advertising in accordance with the method of the present invention;

FIG. 8 is a block diagram that illustrates native code modules for a reception system which might be used for supporting applications and advertising in accordance with the method of the present invention;

FIG. 9 is a schematic diagram that illustrates an example of a partitioned application to be processed by a reception system which might be used for supporting applications and advertising in accordance with the method of the present invention;

FIG. 10 illustrates generation of a page with a page processing table for a reception system which might be used for supporting applications and advertising in accordance with the method of the present invention;

FIG. 11 is a flow diagram for an aspect of the navigation method of a reception system which might be used for supporting applications and advertising in accordance with the method of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

General System Description

FIGS. 1 and 2 show a network in which the method of the present invention for presenting advertising might be used. As seen the network, designated 10, includes a plurality of reception units within a reception layer 401 for displaying information and providing transactional services. In this arrangement, many users each access network 10 with a conventional personal computer; e.g., one of the IBM or IBM-compatible type, which has been provided with application software to constitute a reception system (RS) 400.

US 7,072,849 B1

5

As seen in FIG. 1, interactive network 10 uses a layered structure that includes an information layer 100, a switch/file server layer 200, and cache/concentrator layer 300 as well as reception layer 401. This structure maintains active application databases and delivers requested parts of the databases on demand to the plurality of RS 400's, shown in FIG. 2. As seen in FIG. 2, cache/concentrator layer 300 includes a plurality of cache/concentrator units 302, each or which serve a plurality of RS 400 units over lines 301. Additionally, switch/file server layer 200 is seen to include a server unit 205 connected to multiple cache/concentrator units 302 over lines 201. Still further, server unit 205 is seen to be connected to information layer 100 and its various elements, which act as means for producing, supplying and maintaining the network databases and other information necessary to support network 10. Continuing, switch/filer layer 200 is also seen to include gateway systems 210 connected to server 205. Gateways 210 couple layer 200 to other sources of information and data; e.g., other computer systems. As will be appreciated by those skilled in the art, layer 200, like layers 401 and 300, could also include multiple servers, gateways and information layers in the event even larger numbers of users were sought to be served.

Continuing with reference to FIG. 2, in preferred form, each RS 400 is seen to include a personal computer 405 having a CPU 410 including a microprocessor (as for example, one of the types made by INTEL Corporation in its X'86 family of microprocessors), companion RAM and ROM memory and other associated elements, such as monitor 412 with screen 414 and a keyboard 424. Further, personal computer 405 may also include one or two floppy disk drives 416 for receiving diskettes 426 containing application software used to support the interactive service and facilitate the interactive sessions with network 10. Additionally, personal computer 405 would include operating systems software; e.g., MS-DOS, supplied on diskettes 428 suitable for the personal computer being used. Personal computer 405 still further may also include a hard-disk drive 420 for storing the application software and operating system software which may be transferred from diskettes 426 and 428 respectfully.

Once so configured, each RS 400 provides: a common interface to other elements of interactive computer network 10; a common environment for application processing; and a common protocol for user-application conversation which is independent of the personal computer brand used. RS 400 thus constitutes a universal terminal for which only one version of all applications on network 10 need be prepared, thereby rendering the applications interpretable by a variety of brands of personal computers.

RS 400 formulated in this fashion is capable of communication with the host system to receive information containing either of two types of data, namely objects and messages. Objects have a uniform, self-defining format known to RS 400, and include data types, such as interpretable programs and presentation data for display at monitor screen 414 of the user's personal computer 405. Applications presented at RS 400 are partitioned into objects which represent the minimal units available from the higher levels of interactive network 10 or RS 400. In this arrangement, each application partition typically represents one screen or a partial screen of information, including fields filled with data used in transactions with network 10. Each such screen, commonly called a page, is represented by its parts and is described in a page template object, discussed below.

Applications, having been partitioned into minimal units, are available from higher elements of network 10 or RS 400,

6

and are retrieved on demand by RS 400 for interpretive execution. Thus, not all partitions of a partitioned application need be resident at RS 400 to process a selected partition, thereby raising the storage efficiency of the user's RS 400 and minimizing response time. Each application partition is an independent, self-contained unit and can operate correctly by itself. Each partition may refer to other partitions either statically or dynamically. Static references are built into the partitioned application, while dynamic references are created from the execution of program logic using a set of parameters, such as user demographics or locale. Partitions may be chosen as part of the RS processing in response to user created events, or by selecting a key word of the partitioned application (e.g., "JUMP" or "INDEX," discussed below), which provides random access to all services represented by partitioned applications having key words.

Objects provide a means of packaging and distributing partitioned applications. As noted, objects make up one or more partitioned applications, and are retrieved on demand by a user's RS 400 for interpretive execution and selective storage. All objects are interpreted by RS 400, thereby enabling applications to be developed independently of the personal computer brand used.

Objects may be nested within one another or referenced by an object identifier (object-id) from within their data structure. References to objects permit the size of objects to be minimized. Further, the time required to display a page is minimized when referenced objects are stored locally at RS 400 (which storage is determined by prior usage meeting certain retention criteria), or have been pre-fetched, or in fact, are already used for the current page.

Objects carry application program instructions and/or information for display at monitor screen 414 of RS 400. Application program objects, called pre-processors and post-processors, set up the environment for the user's interaction with network 10 and respond to events created when the user inputs information at keyboard 424 of RS 400. Such events typically trigger a program object to be processed, causing one of the following: sending of transactional information to the coapplications in one layer of the network 10; the receiving of information for use in programs or for presentation in application-dependent fields on monitor screen 414; or the requesting of a new objects to be processed by RS 400. Such objects may be part of the same application or a completely new application.

The RS 400 supports a protocol by which the user and the partitioned applications communicate. All partitioned applications are designed knowing that this protocol will be supported in RS 400. Hence, replication of the protocol in each partitioned application is avoided, thereby minimizing the size of the partitioned application.

RS 400 includes a means to communicate with network 10 to retrieve objects in response to events occurring at RS 400 and to send and receive messages.

RS 400 includes a means to selectively store objects according to a predetermined storage criterion, thus enabling frequently used objects to be stored locally at the RS, and causing infrequently used objects to forfeit their local storage location. The currency of objects stored locally at the RS 400 is verified before use according to the object's storage control parameters and the storage criterion in use for version checking.

Selective storage tailors the contents of the RS 400 memory to contain objects representing all or significant parts of partitioned applications favored by the user.

US 7,072,849 B1

7

Because selective storage of objects is local, response time is reduced for those partitioned applications that the user accesses most frequently.

Since much of the application processing formerly done by a host computer in previously known time-sharing networks is now performed at the user's RS 400, the higher elements of network 10, particularly layer 200, have as their primary functions the routing of messages, serving of objects, and line concentration. The narrowed functional load of the higher network elements permits many more users to be serviced within the same bounds of computer power and I/O capability of conventional host-centered architectures.

Network 10 provides information on a wide variety of topics, including, but not limited to news, industry, financial needs, hobbies and cultural interests. Network 10 thus eliminates the need to consult multiple information sources, giving users an efficient and timesaving overview of subjects that interest them.

The transactional features of interactive network 10 saves the user time, money, and frustration by reducing time spent traveling, standing in line, and communicating with sales personnel. The user may, through RS 400, bank, send and receive messages, review advertising provided in accordance with the method of the present invention, place orders for merchandise, and perform other transactions.

In preferred form, network 10 provides information, advertising and transaction processing services for a large number of users simultaneously accessing the network via the public switched telephone network (PSTN), broadcast, and/or other media with their RS 400 units. Services available to the user include display of information such as movie reviews, the latest news, airlines reservations, the purchase of items such as retail merchandise and groceries, and quotes and buy/sell orders for stocks and bonds. Network 10 provides an environment in which a user, via RS 400 establishes a session with the network and accesses a large number of services. These services are specifically constructed applications which as noted are partitioned so they may be distributed without undue transmission time, and may be processed and selectively stored on a user's RS 400 unit.

System Configuration

As shown in FIG. 1, interactive computer network 10 includes four layers: information layer 100, switch and file server layer 200, concentrator layer 300, and reception layer 401.

Information layer 100 handles: (1) the production, storage and dissemination of data and (2) the collection and off-line processing of such data from each RS session with the network 10 so as to permit the targeting of information and advertising to be presented to users and for traditional business support.

Switch and file server layer 200 and cache/concentrator layer 300 together constitute a delivery system 20 which delivers requested data to the RS 400's of reception layer 401 and routes data entered by the user or collected at RS 400's to the proper application in network 10. With reference to FIG. 2, the information used in a RS 400 either resides locally at the RS 400, or is available on demand from the cache/concentrator 300 or the file server 205, via the gateway 210, which may be coupled to external providers, or is available from information layer 100.

There are two types of information in the network 10 which are utilized by the RS 400: objects and messages.

8

Objects include the information requested and utilized by the RS 400 to permit a user to select specific parts of applications, control the flow of information relating to the applications, and to supply information to the network. Objects are self-describing structures organized in accordance with a specific data object architecture, described below. Objects are used to package presentation data and program instructions required to support the partitioned applications and advertising presented at a RS 400. Objects are distributed on demand throughout interactive network 10. Objects may contain: control information; program instructions to set up an application processing environment and to process user or network created events; information about what is to be displayed and how it is to be displayed; references to programs to be interpretively executed; and references to other objects, which may be called based upon certain conditions or the occurrence of certain events at the user's personal computer, resulting in the selection and retrieval of other partitioned applications packaged as objects.

Messages are information provided by the user or the network and are used in fields defined within the constructs of an object, and are seen on the user's RS monitor 412, or are used for data processing at RS 400. Additionally, and as more fully described hereafter, messages are the primary means for communication within and without the network. The format of messages is application dependent. If the message is input by the user, it is formatted by the partitioned application currently being processed on RS 400. Likewise, and with reference to FIG. 2, if the data are provided from a co-application database residing in delivery system 20, or accessed via gateway 210 or high function system 110 within the information layer 100, the partitioned application currently being processed on RS 400 causes the message data to be displayed in fields on the user's display monitor as defined by the particular partitioned application.

All active objects reside in file server 205. Inactive objects or objects in preparation reside in producer system 120. Objects recently introduced into delivery system 20 from the producer system 120 will be available from file server 205, but, may not be available on cache/concentrator 302 to which the user's RS 400 has dialed. If such objects are requested by the RS 400, the cache/concentrator 302 automatically requests the object from file server 205. The requested object is routed back to the requesting cache/concentrator 302, which automatically routes it to the communications line on which the request was originally made, from which it is received by the RS 400.

The RS 400 is the point of application session control because it has the ability to select and randomly access objects representing all or part of partitioned applications and their data. RS 400 processes objects according to information contained therein and events created by the user on personal computer 405.

Applications on network 10 act in concert with the distributed partitioned applications running on RS 400. Partitioned applications constructed as groups of objects and are distributed on demand to a user's RS 400. An application partition represents the minimum amount of information and program logic needed to present a page or window, i.e. portion of a page presented to the user, perform transactions with the interactive network 10, and perform traditional data processing operations, as required, including selecting another partitioned application to be processed upon a user generated completion event for the current partitioned application.

US 7,072,849 B1

9

Objects representing all or part of partitioned applications may be stored in a user's RS 400 if the objects meet certain criteria, such as being non-volatile, non-critical to network integrity, or if they are critical to ensuring reasonable response time. Such objects are either provided on diskettes 426 together with RS 400 system software used during the installation procedure or they are automatically requested by RS 400 when the user makes selections requiring objects not present in RS 400. In the latter case, RS 400 requests from cache/concentrator layer 300 only the objects necessary to execute the desired partitioned application.

Reception system application software 426 in preferred form is provided for IBM and IBM-compatible brands of personal computers 405, and all partitioned applications are constructed according to a single architecture which each such RS 400 supports. With reference to FIG. 2, to access network 10, a user preferably has a personal computer 405 with at least 512K RAM and a single disk drive 416. The user typically accesses network 10 using a 1,200 or 2,400 bps modem (not shown). To initiate a session with network 10, objects representing the logon application are retrieved from the user's personal diskette, including the R.S. application software, which was previously set up during standard installation and enrollment procedures with network 10. Once communication between RS 400 and cache/concentrator layer 300 has been established, the user begins a standard logon procedure by inputting a personal entry code. Once the logon procedure is complete, the user can begin to access various desired services (i.e., partitioned applications) which provide display of requested information and/or transaction operations.

Applications and Pages

Applications, i.e. information events, are composed of a sequence of one or more pages opened at screen 414 of monitor 412. This is better seen with reference to FIGS. 3a and 3b were a page 255 is illustrated as might appear at screen 414 of monitor 412. With reference to FIG. 3a, each page 255 is formatted with a service interface having page partitions 250, 260, 280, and 290 (not to be confused with application partitions). Window page partitions 275, well known in the art, are also available and are opened and closed conditionally on page 255 upon the occurrence of an event specified in the application being run. Each page partition 250, 260, 280 and 290 and window 275 is made up of a page element which defines the content of the partition or window.

Each page 255 includes: a header page partition 250, which has a page element associated with it and which typically conveys information on the page's topic or sponsor; one or more body page partitions 260 and window page partitions 275, each of which is associated with a page element which as noted gives the informational and transactional content of the page. For example, a page element may contain presentation data selected as a menu option in the previous page, and/or may contain prompts to which a user responds in pre-defined fields to execute transactions. As illustrated in FIG. 3b, the page element associated with body page partition 260—includes display fields 270, 271, 272. A window page partition 275 seen in FIG. 3a represents the same informational and transactional capability as a body partition, except greater flexibility is provided for its location and size.

Continuing with reference to FIG. 3a, in accordance with the invention, advertising 280 is provided over network 10, like page elements, also includes information for display on

10

page 255, and may be included in any partition of a page. Advertising 280 is presented to the user on an individualized basis from queues of advertising object identifications (ids) that are constructed off-line by business system 130, and sent to file server 205 where they are accessible to each RS 400.

Individualized queues of advertising object ids are constructed based upon data collected on the partitioned applications that were accessed by a user, and upon events the user generated in response to applications. The data are collected and reported by RS 400 to a data collection co-application in file server 205 for later transmission to business system 130. In addition to application access and use characteristics, a variety of other parameters, such as user demographics or postal ZIP code, may be used as targeting criteria. From such data, queues of advertising object ids are constructed that are targeted to either individual users or to sets of users who fall into certain groups according to such parameters. Stated otherwise, the advertising presented is individualized to the respective users based on characterizations of the respective users as defined by the interaction history with the service and such other information as user demographics and locale. As will be appreciated by those skilled in the art, conventional marketing analysis techniques can be employed to establish the user characterizations based on the collected application usage data above noted and other information.

Also with reference to FIG. 3b, the service interface is seen to include a command region 285 which enables the user to interact with the network RS 400 and other elements of network 10, so as to cause such operations as navigating from page to page, performing a transaction, or obtaining more information about other applications. As shown in FIG. 3b, interface region 285 includes a command bar 290—having a number of commands 291–298 which the user can execute. The functions of commands 291–298 are discussed in greater detail below.

Network Objects

As noted above, in conventional time-sharing computer networks, the data and program instructions necessary to support user sessions are maintained at a central host computer. However, that approach has been found to create processing bottlenecks as greater numbers of users are connected to the network; bottlenecks which require increases in processing power and complexity; e.g., multiple hosts of greater computing capability, if the network is to meet demand. Further, such bottlenecks have been found to also slow response time as more users are connected to the network and seek to have their requests for data processing answered.

The consequences of the host processing bottlenecking is to either compel capital expenditures to expand host processing capability, or accept longer response times; i.e., a slower network, and risk user dissatisfaction.

However, even in the case where additional computing power is added, and where response time is allowed to increase, eventually the host becomes user saturated as more and more users are sought to be served by the network. The network described above, however, is designed to alleviate the effects of host-centered limitations, and extend the network saturation point. This objective is achieved by reducing the demand on the host for processing resources by structuring the network so that the higher network levels act primarily to maintain and supply data and programs to the

US 7,072,849 B1

11

lower levels of the network, particularly RS 400, which acts to manage and sustain the user screen displays.

More particularly, the described network features procedures for parsing the network data and program instructions required to support the interactive user sessions into packets, referred to as objects, and distributing them into the network where they can be processed at lower levels, particularly, reception system 400.

In accordance with the method of the present invention, the screens presented at the user's monitor are each divided into addressable partitions shown in FIG. 3a, and the display text and graphics necessary to make up the partitions, as well as the program instructions and control data necessary to deliver and sustain the screens and partitions, are formulated from pre-created objects. Further, the objects are structured in accordance with an architecture that permits the displayed data to be relocatable on the screen, and to be reusable to make up other screens and other sessions, either as pre-created and stored sessions or interactive sessions, dynamically created in response to the user's requests.

As shown in FIG. 4c, the network objects are organized as a family of objects each of which perform a specific function in support of the interactive session. More particularly, in accordance with the preferred form of the invention, the network object family is seen to include 6 members: page format objects 502, page element objects 504, window objects 506, program objects 508, advertisement objects 510 and page template objects 500.

Within this family, page format objects 502 are designed to define the partitioning 250 to 290 of the monitor screen shown in FIG. 3a. The page format objects 502 provide a means for pre-defining screen partitions and for ensuring a uniform look to the page presented on the reception system monitor. They provide the origin; i.e., drawing points, and dimensions of each page partition and different values for presentation commands such as palette and background color.

Page format objects 502 are referenced whenever non-window data is to be displayed and as noted ensure a consistent presentation of the page. In addition, page format objects 502 assures proper tessellation or "tiling" of the displayed partitions.

Page element objects 504, on the other hand, are structured to contain the display data; i.e., text and graphic, to be displayed which is mapped within screen partitions 250 to 290, and to further provide the associated control data and programs. More specifically, the display data is described within the object as NAPLPS data, and includes, PDI, ASCII, Incremental Point and other display encoding schemes. Page element objects also control the functionality within the screen partition by means of field definition segments 516 and program call segments 532, as further described in connection with the description of such segments hereafter. Page element objects 504 are relocatable and may be reused by many pages. To enable the displayable data to be relocated, display data must be created by producers in the NAPLPS relative mode.

Continuing with reference to FIG. 4c, window objects 506 include the display and control data necessary to support window partitions 275 best seen in FIG. 3a. Windows contain display data which overlay the base page and control data which supersedes the base page control data for the underlying screen during the duration of the window. Window objects 506 contain data which is to be displayed or otherwise presented to the viewer which is relatively independent from the rest of the page. Display data within windows overlay the base page until the window is closed.

12

Logic associated with the window supersedes base page logic for the duration of the window. When a window is opened, the bit map of the area covered by window is saved and most logic functions for the overlaid page are deactivated. When the window is closed, the saved bit map is swapped onto the screen, the logic functions associated with the window are disabled, and prior logic functions are reactivated.

Windows are opened by user or program control. They do not form part of the base page. Windows would typically be opened as a result of the completion of events specified in program call segments 532.

Window objects 506 are very similar in structure to page element objects 504. The critical difference is that window objects 506 specify their own size and absolute screen location by means of a partition definition segment 528.

Program objects 508 contain program instructions written in a high-level language called TRINTEX Basic Object Language, i.e., TBOL, described in greater detail hereafter, which may be executed on RS 400 to support the application. More particularly, program objects 508 include interpretable program code, executable machine code and parameters to be acted upon in conjunction with the presentation of text and graphics to the reception system monitors.

Program objects 508 may be called for execution by means of program call segments 532, which specify when a program is to be executed (event), what program to execute (program pointer), and how programs should run (parameters).

Programs are treated as objects to conform to the open-ended design philosophy of the data object architecture (DOA), allowing the dissemination of newly developed programs to be easily and economically performed. As noted above, it is desirable to have as many of these program objects staged for execution at or as close to RS 400 as possible.

Still further, in accordance with the method of the present invention, advertising objects 510 include the text and graphics that may be presented at ad partition 280 presented on the monitor screen as shown in FIG. 3b.

Finally, the object family includes page template objects 500. Page template objects 500 are designed to define the components of the full screen presented to the viewer. Particularly, page template objects 500 include the entry point to a screen, the name of the page format objects which specify the various partitions a screen will have and the page element object that contain the display data and partitioning parameters for the page.

Additionally, page template object 500 includes the specific program calls required to execute the screens associated with the application being presented to the user, and may serve as the means for the user to selectively move through; i.e., navigate the pages of interest which are associated with various applications. Thus, in effect, page template objects 500 constitute the "recipe" for making up the collection of text and graphic information required to make the screens to be presented to the user.

Also in accordance with the invention, object 500 to 510 shown in FIG. 4c are themselves made up of further sub-blocks of information that may be selectively collected to define the objects and resulting pages that ultimately constitute the application presented to the user in an interactive text and graphic session.

More specifically and as shown schematically in FIG. 4a, objects 500 to 510 are predefined, variable length records consisting of a fixed length header 551 and one or more

US 7,072,849 B1

13

self-defining record segments **552** a list of which is presented in FIG. 4c as segment types **512** to **540**.

In accordance with this design, and as shown in FIG. 4b, object header **551** in preferred form is 18 bytes in length and contains a prescribed sequence of information which provides data regarding the object's identification, its anticipated use, association to other objects, its length and its version and currency.

More particularly, each of the 18 bytes of object header **551** are conventional hexadecimal, 8 bit bytes and are arranged in a fixed pattern to facilitate interpretation by network **10**. Particularly, and as shown in FIG. 4b, the first byte of header **551**; i.e., byte **1**, identifies the length of the object ID in hexadecimal. The next six bytes; i.e., bytes **2** to **7**, are allocated for identifying access control to the object so as to allow creation of closed user groups to whom the object(s) is to be provided. As will be appreciated by those skilled in the art, the ability to earmark objects in anticipation of user requests enables the network anticipate requests and pre-collect objects from large numbers of them maintained to render the network more efficient and reduce response time. The following 4 bytes of header **551**; bytes **8** to **11**, are used to identify the set of objects to which the subject object belongs. In this regard, it will be appreciated that, again, for speed of access and efficiency of selection, the objects are arranged in groups or sets which are likely to be presented to user sequentially in presenting the page sets; i.e., screens that go to make up a session.

Following identification of the object set, the next byte in header **551**; i.e., byte **12**, gives the location of the subject object in the set. As will be appreciated here also the identification is provided to facilitate ease of object location and access among the many thousands of objects that are maintained to, thereby, render their selection and presentation more efficient and speedy.

Thereafter, the following bytes of header **551**; i.e., byte **13**, designates the object type; e.g., page format, page template, page element, etc. Following identification of the object type, two bytes; i.e., bytes **14**, **15**, are allocated to define the length of the object, which may be of whatever length is necessary to supply the data necessary, and thereby provides great flexibility for creation of the screens. Thereafter, a single byte; i.e., byte **16**, is allocated to identify the storage characteristic for the object; i.e., the criterion which establishes at what level in network **10** the object will be stored, and the basis upon which it will be updated. At least a portion of this byte; i.e., the higher order nibble (first 4 bits reading from left to right) is associated with the last byte; i.e., byte **18**, in the header which identifies the version of the object, a control used in determining how often in a predetermined period of time the object will be updated by the network.

Following storage characteristic byte **16**, header **551** includes a byte; i.e., **17**, which identifies the number of objects in the set to which the subject object belongs. Finally, and as noted above, header **551** includes a byte; i.e., **18**, which identifies the version of the object. Particularly the object version is a number to establish the control for the update of the object that are resident at RS **400**.

As shown in FIG. 4a, and as noted above, in addition to header **551**, the object includes one more of the various segment types shown in FIG. 4c.

Segments **512** to **540** are the basic building blocks of the objects. And, as in the case of the object, the segments are also self-defining. As will be appreciated by those skilled in the art, by making the segments self-defining, changes in the

14

objects and their use in the network can be made without changing pre-existing objects.

As in the case of objects, the segments have also been provided with a specific structure. Particularly, and as shown in FIG. 4a, segments **552** consists of a designation of segment type **553**, identification of segment length **554**, followed by the information necessary to implement the segment and its associated object **555**; e.g., either, control data, display data or program code.

10 In this structure, segment type **553** is identified with a one-byte hexadecimal code which describes the general function of the segment. Thereafter, segment length **554** is identified as a fixed two-byte long field which carries the segment length as a hexadecimal number in INTEL format; i.e., least significant byte first. Finally, data within segments may be identified either by position or keyword, depending on the specific requirements of the segment.

The specific structure for the objects and segments is shown in FIG. 4c and is described below. In that description the following notation convention is used:

<>—mandatory item

()—optional item

...—item may be repeated

|item| |item|

<> ()—items in a column indicate either/or

|item| |item|

The structure for objects is:

PAGE TEMPLATE OBJECT,

30 [<header> (compression descriptor) <page format call> (page element call) ... (program call) ... (page element selector) (system table call) ... external reference) (keyword/navigation) ...];

35 As noted above, page format objects **502** are designed to define the partitioning **250** to **290** of monitor screen **414** shown in FIG. 3a.

PAGE FORMAT OBJECT,

40 [<header> (compression descriptor) (page defaults) <partition definition>];

PAGE ELEMENT OBJECT,

45 [<header> (compression descriptor) (presentation data) ... (program call) ... (custom cursor) ... (custom text) ... (field definition) ... (field-level program call) ... (custom cursor type 2) ... (custom graphic) ... (field definition type 2) ... (array definition) ... (inventory control)];

50 Page element objects, as explained, are structured to contain the display data; i.e., text and graphics, to be presented at screen partitions **250** to **290**.

WINDOW OBJECT,

55 [<header> (compression description) <partition definition> (page element call) (presentation data) ... (program call) ... (custom cursor) ... (custom text) ... (custom cursor type 2) ... (custom graphic) ... (field definition) ... (field level program call) ... (field definition type 2) ... (array definition) ... (inventory control)];

60 As noted, window objects include display and control data necessary to support window partition at screen **414**.

PROGRAM OBJECTS,

65 [<header> (compression descriptor) <program data> ...].

Program objects, on the other hand, contain program instructions written in higher-level language which may be executed at RS **400** to support the application.

US 7,072,849 B1

15

ADVERTISEMENT OBJECT,

[<header> (compression descriptor) (presentation data) . . . (program call) . . . (custom cursor) . . . (custom text) . . . (field definition) . . . (field-level program call) . . . (custom cursor type 2) . . . (custom graphic) . . . (field definition type 2) . . . (array definition) . . . (inventory control)];

In accordance with the invention, and as can be seen, advertisement objects are substantially the same as page element objects, with the difference being that, as their name implies, their subject matter is selected to concern advertising.

Continuing, the structure for the object segments follows from the above description, and is as described more fully in parent application Ser. No. 388,156 now issued as U.S. Pat. No. 5,347,632, the contents of which patent are incorporated herein by reference.

Network Messages

In addition to the network objects, and the display data, control data, and the program instructions they contain as previously described, network **10** also exchanges information regarding the support of user sessions and the maintenance of the network as "messenger". Specifically, messages typically relate to the exchange of information associated with initial logon of a reception system **400** to network **10**, dialogue between RS **400** and other elements and communications by the other network elements amongst themselves.

To facilitate message exchange internally, and through gateway **210** to entities externally to network **10**, a protocol termed the "Data Interchange Architecture" (DIA) is used to support the transport and interpretation of information. More particularly, DIA enables: communications between RS **400** units, separation of functions between network layers **100**, **200**, **300** and **401**; consistent parsing of data; an "open" architecture for network **10**; downward compatibility within the network; compatibility with standard industry protocols such as the IBM System Network Architecture; Open Systems Interconnections standard; support of network utility sessions; and standardization of common network and application return codes.

Thus DIA binds the various components of network **10** into a coherent entity by providing a common data stream for communications management purposes. DIA provides the ability to route messages between applications based in IBM System Network Architecture (SNA), (well known in the art, and more fully described in *Data and Computer Communications*, by W. Stallings, Chapter 12, McMillian Publishing, Inc. (1985)) and non-SNA reception system applications; e.g. home computer applications. Further, DIA provides common data structure between applications run at RS **400** units and applications that may be run on external computer networks; e.g. Dow Jones Services, accessed through gateway **210**. As well, DIA provides support for utility sessions between backbone applications run within network **10**. A more detailed description of network messaging is provided in above noted U.S. Pat. No. 5,347,632, the content of which is incorporated herein by reference.

Object Language

In accordance with the design of network **10**, in order to enable the manipulation of the network objects, the application programs necessary to support the interactive text/graphic sessions are written in a high-level language referred

16

to as "TBOL", (TRINTEX Basic Object Language, "TRINTEX" being the former company name of one of the assignees of this invention). TBOL is specifically adapted for writing the application programs so that the programs may be compiled into a compact data stream that can be interpreted by the application software operating in the user personal computer, the application software being designed to establish the network Reception System **400** previously noted and described in more detail hereafter.

The Reception System application software supports an interactive text/graphics sessions by managing objects. As explained above, objects specify the format and provide the content; i.e., the text and graphics, displayed on the user's screen so as to make up the pages that constitute the application. As also explained, pages are divided into separate areas called "partitions" by certain objects, while certain other objects describe windows which can be opened on the pages. Further, still other objects contain TBOL application programs which facilitate the data processing necessary to present the pages and their associated text and graphics.

As noted, the object architecture allows logical events to be specified in the object definitions. An example of a logical event is the completion of data entry on a screen; i.e., an application page. Logical events are mapped to physical events such as the user pressing the <ENTER> key on the keyboard. Other logical events might be the initial display of a screen page or the completion of data entry in a field. Logical events specified in page and window object definitions can be associated with the call of TBOL program objects.

RS **400** is aware of the occurrence of all physical events during the interactive text/graphic sessions. When a physical event such as depression of the forward <TAB> key corresponds to a logical event such as completion of data entry in a field, the appropriate TBOL program is executed if specified in the object definition. Accordingly, the TBOL programs can be thought of as routines which are given control to perform initialization and post-processing application logic associated with the fields, partitions and screens at the text/graphic sessions.

RS **400** run time environment uses the TBOL programs and their high-level key-word commands called verbs to provide all the system services needed to support a text/graphic session, particularly, display management, user input, local and remote data access.

TBOL programs have a structure that includes three sections: a header section in which the program name is specified; a data section in which the data structure the program will use are defined; and a code section in which the program logic is provided composed of one or more procedures. More specifically, the code section procedures are composed of procedure statements, each of which begins with a TBOL key word called a verb.

The name of a procedure can also be used as the verb in a procedure statement exactly as if it were a TBOL key-word verb. This feature enables a programmer to extend the language vocabulary to include customized application-oriented verb commands.

Continuing, TBOL programs have a program syntax that includes a series of "identifiers" which are the names and labels assigned to programs, procedures, and data structures.

An identifier may be up to 31 characters long; contain only uppercase or lowercase letters A through Z, digits 0 through 9, and/or the special character underscore (); and must begin with a letter. Included among the system identifiers are: "header section identifiers" used in the header section for the program name; "data section identifiers" used

US 7,072,849 B1

17

in the data section for data structure names, field names and array names; and finally, "code section identifiers" used in the code section for identification of procedure names and statement labels. A more detailed description of TBOL is provided in parent application Ser. No. 388,156 now issued as U.S. Pat. No. 5,347,632, the contents of which patent are incorporated herein by reference.

Reception System Operation

RS 400 of computer system network 10 uses software called native code modules (described below) to enable the user to select options and functions presented on the monitor screen 414 of personal computer 405, to execute partitioned applications and to process user created events, enabling the partitioned application to interact with network 10. Through this interaction, the user is able to input data into fields provided as part of the display, or may individually select choices causing a standard or personalized page to be built (as explained below) for display on the monitor of personal computer 405. Such inputs will cause RS 400 to interpret events and trigger pre-processors or post-processors, retrieve specified objects, communicate with system components, control user options, cause the display of advertisements on a page, open or close window partitions to provide additional navigation possibilities, and collect and report data about events, including certain types of objects processed. For example, the user may select a particular option, such as opening or closing window partition 275, which is present on the monitor and follow the selection with a completion key stroke, such as ENTER. When the completion keystroke is made, the selection is translated into a logical event that triggers the execution of a post-processor, (i.e., a partitioned application program object) to process the contents of the field.

Functions supporting the user-partitioned application interface can be performed using the command bar 290, or its equivalent using pull down windows or an overlapping cascade of windows. These functions can be implemented as part of the RS native functions or can be treated as another partition(s) defined for every page for which an appropriate set of supporting objects exist and remain resident at RS 400. If the functions are part of RS 400, they can be altered or extended by verbs defined in the RS virtual machine that permit the execution of program objects to be triggered when certain functions are called, providing maximum flexibility.

To explain the functions the use of a command bar is assumed. Command bar 290 is shown in FIGS. 3a and 3b and includes a NEXT command 291, a BACK command 292, a PATH command 293, a MENU command 294, an ACTION command 295, a JUMP command 296, a HELP command 297, and an EXIT command 298.

NEXT command 291 causes the next page in the current page set to be built. If the last page of a page set has already been reached, NEXT command 291 is disabled by RS 400, avoiding the presentation of an invalid option.

BACK command 292 causes the previous page of the current page set to be built. If the present page is the first in the page set, BACK command 292 is disabled, since it is not a valid option.

A filter program can be attached to both the NEXT or BACK functions to modify their implicit sequential nature based upon the value of the occurrence in the object set id.

PATH command 293 causes the next page to be built and displayed from a list of pages that the user has entered, starting from the first entry for every new session.

18

MENU command 294 causes the page presenting the previous set of choices to be rebuilt.

ACTION command 295 initiates an application dependent operation such as causing a new application partition to be interpreted, a window partition 275 to be opened and enables the user to input any information required which may result in a transaction or selection of another window or page.

JUMP command 296 causes window partition 275 to be opened, allowing the user to input a keyword or to specify one from an index that may be selected for display.

HELP command 297 causes a new application partition to be interpreted such as a HELP window pertaining to where the cursor is positioned to be displayed in order to assist the user regarding the present page, a particular partition, or a field in a page element.

EXIT command 298 causes a LOGOFF page template object (PTO) to be built, and a page logoff sequence to be presented at RS 400 monitor screen 414.

Navigation Interface

Continuing, as a further feature, network 10 includes an improved procedure for searching and retrieving applications from the store of applications distributed throughout network 10; e.g., server 205, cache/concentrator 302 and RS 400. More specifically, the procedure features use of pre-created search tables which represent subsets of the information on the network arranged with reference to the page template objects (PTO) and object-ids of the available applications so that in accordance with the procedure, the relevant tables and associated objects can be provided to and searched at the requesting RS 400 without need to search the entire store of applications on the network. As will be appreciated, this reduces the demand on the server 205 for locating and retrieving applications for display at monitor 412.

In conventional time-sharing networks that support large conventional databases, the host receives user requests for data records; locates them; and transmits them back to the users. Accordingly, the host is obliged to undertake the data processing necessary to isolate and supply the requested information. And, as noted earlier, where large numbers of users are to be served, the many user requests can bottleneck at the host, taxing resources and leading to response slowdown.

Further, users have experienced difficulty in searching data bases maintained on conventional time-sharing networks. For example, difficulties have resulted from the complex and varied way previously known database suppliers have organized and presented their information. Particularly, some database providers require searching be done only in selected fields of the data base, thus requiring the user to be fully familiar with the record structure. Others have organized their databases on hierachial structures which require the user understand the way the records are grouped. Still further, yet other database suppliers rely upon keyword indices to facilitate searching of their records, thus requiring the user to be knowledgeable regarding the particular keywords used by the database provider.

Network 10, however, is designed to avoid such difficulties. In the preferred embodiment, the network includes procedures for creating preliminary searches which represent subsets of the network applications users are believed likely to investigate. Particularly, in accordance with these procedures, for the active applications available on network 10, a library of tables is prepared, and maintained within

US 7,072,849 B1

19

each of which a plurality of so called "keywords" are provided that are correlated with page template objects and object-ids of the entry screen (typically the first screen) for the respective application. In the preferred embodiment, approximately 1,000 tables are used, each having approximately 10 to 20 keywords arranged in alphabetical order to abstract the applications on the network. Further, the object-id for each table is associated with a code in the form of a character string mnemonic which is arranged in a set of alphabetically sequenced mnemonics termed the sequence set so that on entry of a character string at an RS 400, the object-id for the relevant keyword table can be obtained from the sequence set. Once the table object-id is identified, the keyword table corresponding to the desired subset of the objects and associated applications can then be obtained from network 10. Subsequently the table can be presented to the user's RS 400, where the RS 400 can provide the data processing required to present the potentially relevant keywords, objects and associated applications to the user for further review and determination as to whether more searching is required. As will be appreciated, this procedure reduces demand on server 205 and thereby permits it to be less complex and costly, and further, reduces the likelihood of host overtaxing that may cause network response slowdown.

As a further feature of this procedure, the library of keywords and their associated PTOs and objects may be generated by a plurality of operations which appear at the user's screen as different search techniques. This permits the user to select a search technique he is most comfortable with, thus expediting his inquiry.

More particularly, the user is allowed to invoke the procedure by calling up a variety of operations. The various operations have different names and seemingly present different search strategies. Specifically, the user may invoke the procedure by initiating a "Jump" command at RS 400. Thereafter, in connection with the Jump operation, the user, when prompted, may enter a word of the user's choosing at monitor screen 414 relating to the matter he is interested in locating; i.e., a subject matter search of the network applications. Additionally, the users may invoke the procedure by alternatively calling up an operation termed "Index" with selection of the Index command. When selected, the Index command presents the user with an alphabetical listing of keywords from the tables noted above which the user can select from; i.e., an alphabetical search of the network applications. Further, the user may evoke the procedure by initiating an operation termed "Guide." By selecting the Guide command, the user is provided with a series of graphic displays that presents a physical description of the network applications; e.g., department floor plan for a store the user may be electronically shopping in. Still further, the user may invoke the procedures by initiating an operation termed "Directory." By selecting the Directory command, the user is presented with the applications available on the network as a series of hierarchical menus which present the content of the network information in commonly understood categories. Finally, the user may invoke the procedure by selecting the "Path" command, which accesses a list of keywords the user has previously selected; i.e., a personally tailored form of the Index command described above. As described hereafter, Path further includes a Viewpath operation which permits the user to visually access and manage the Path list of keywords. In preferred form, where the user has not selected a list of personalized keywords, a default set

20

is provided which includes a predetermined list and associated applications deemed by network 10 as likely to be of interest to the user.

This ability to convert these apparently different search strategies in a single procedure for accessing pre-created library tables is accomplished by translating the procedural elements of the different search techniques into a single set of procedures that will produce a mnemonic; i.e., code word, which can first be searched at the sequence set, described above to identify the object-id for the appropriate library table and, thereafter, enable access of the appropriate table to permit selection of the desired keyword and associated PTO and object-ids. That is to say, the reception system native code simply relates the user-entered character string, alphabetical range, category, or list item of respectively, "Jump", "Index", "Directory", or "Path" to the table codes through the sequence set, so that the appropriate table can be provided to the reception system and application keyword selected. Thus, while the search techniques may appear different to the user, and in fact accommodate the user's preferences and sophistication level, they nonetheless invoke the same efficient procedure of relying upon pre-created searches which identify related application PTOs and object-ids so that the table and objects may be collected and presented at the user's RS 400 where they can be processed, thereby relieving server 205.

In preferred form, however, in order to enhance presentation speed the Guide operation is specially configured. Rather than relating the keyword mnemonic to a sequence set to identify the table object-id and range of keywords corresponding to the entry PTO and associated object-ids, the Guide operation presents a series of overlapping windows that physically describe the "store" in which shopping is being conducted or the "building" from which information is being provided. The successive windows increase in degree of detail, with the final window presenting a listing of relevant keywords. Further, the PTO and object-ids for the application entry screen are directly related to the graphic presentation of the keywords. This eliminates the need to provide variable fields in the windows for each of the keywords and enables the entry screen to be correlated directly with the window graphic. As will be appreciated, this reduces the number of objects that would otherwise be required to be staged at RS 400 to support pretention of the keyword listing at monitor screen 414, and thus speeds network response.

A more detailed understanding of the procedure may be had upon a reading of the following description and review of accompanying FIGS. 2, 3a and particularly FIG. 11 which presents a flow diagram for the Jump sequence of the search procedure.

To select a particular partitioned application from among thousands of such applications residing either at the RS 400 or within delivery system 20, network 10 avoids the need for a user to know or understand, prior to a search, the organization of such partitioned applications and the query techniques necessary to access them. This is accomplished using a collection of related commands, as described below.

The Jump command 296 as seen in FIG. 3a, can be selected, by the user from command bar 290. When Jump command 296 is selected, a window partition 275 is opened. In window 275, the user is presented and may select from a variety of displayed options that include among others, the Directory command, the Index command, and the Guide command, which when selected, have the effect noted above. Additionally, the user can select a command termed Viewpath which will presents the keywords that currently

US 7,072,849 B1

21

make up the list of keywords associated with the user's Path command, and from which list the user can select a desired keyword. Still further, and with reference FIG. 11, which shows the sequence where a user offers a term to identify a subject of interest, the user may enter a keyword at display field 270 within window partition 275 as a "best guess" of the mnemonic character string that is assigned to a partitioned application the user desires (e.g., the user may input such english words as "news," "pet food," "games," etcetera). Where the user enters a character string it is displayed in field 270, and then searched by RS 400 native code (discussed below) against the sequence sets above noted to identify the object-id for the appropriate table of keywords (not shown) that RS 400 may request from host 205. While as noted above, a table may include 10 to 20 keywords, in the preferred embodiment, for the sake of speed and convenience, a typical keyword table includes approximately 12 keywords.

If the string entered by the user matches a keyword existing on one of the keyword tables, and is thus associated with a specific PTO, RS 400 fetches and displays associated objects of the partitioned applications and builds the entry page in accordance with the page composition dictated by the target PTO.

If the string entered by the user does not match a specific keyword, RS 400 presents the user with the option of displaying the table of keywords approximating the specific keyword. The approximate keywords are presented as initialized, cursorable selector fields of the type provided in connection with a Index command. The user may then move the cursor to the nearest approximation of the mnemonic he originally selected, and trigger navigation to the PTO associated with that keyword, navigation being as described hereafter in connection with the RS 400 native code.

If, after selecting the Jump command, the user selects the Index command, RS 400 will retrieve the keyword table residing at RS 400, and will again build a page with initialized, cursorable fields of keywords. The table fetched upon invoking the Index command will be comprised of alphabetic keywords that occur within the range of the keywords associated with the page template object (PTO) from which the user invoked the Index command. As discussed above, the user may select to navigate to any of this range of PTOs by selecting the relevant keyword from the display. Alternatively, the user can, thereafter, select another range of alphabetical keywords by entering an appropriate character string in a screen field provided or move forward or backward in the collection by selecting the corresponding option.

By selecting the Directory command, RS 400 can be caused to fetch a table of keywords, grouped by categories, to which the PTO of the current partitioned application (as specified by the object set field 630 of the current PEO) belongs. Particularly, by selecting the Directory command, RS 400, is causes to displays a series of screens each of which contains alphabetically arranged general subject categories from which the user may select. Following selection of a category, a series of keywords associated with the specified category are displayed in further screens together with descriptive statements about the application associated with the keywords. Thereafter, the user can, in the manner previously discussed with regard to the Index command, select from and navigate to the PTOs of keywords which are related to the present page set by subject.

The Guide command provides a navigation method related to a hierarchical organization of applications provided on network 10, and are described by a series of

22

sequentially presented overlaying windows of a type known in the art, each of which presents an increasing degree of detail for a particular subject area, terminating in a final window that gives keywords associated with the relevant applications. The Guide command makes use of the keyword segment which describes the location of the PTO in a hierarchy (referred to, in the preferred embodiment, as the "BFD," or Building-Floor-Department) as well as an associated keyword character string. The BFD describes the set of menus that are to be displayed on the screen as the sequence of pop-up windows. The Guide command may be invoked by requesting it from the Jump window described above, or by selecting the Menu command on Command Bar 290. As noted above, in the case of the Guide command, the PTO and object-ids for the application entry screen are directly associated with the graphic of the keyword presented in the final pop-up window. This enables direct access of the application entry screen without need to access the sequence set and keyword table, and thus, reduces response time by reducing the number of objects that must be processed at RS 400.

Activation of the Path command accesses the user's list of pre-selected keywords without their display, and permits the user to step through the list viewing the respective applications by repeatedly invoking the Path command. As will be appreciated, the user can set a priority for selecting keywords and viewing their associated applications by virtue of where on the list the user places the keywords. More specifically, if the user has several application of particular interest; e.g., news, weather, etc., the user can place them at the top of the list, and quickly step through them with the Path command. Further, the user can view and randomly access the keywords of his list with the Viewpath operation noted above. On activation of Viewpath, the user's Path keywords are displayed and the user can cursor through them in a conventional manner to select a desired one. Further, the user can amend the list as desired by changing the keywords on the list and/or adjusting their relative position. This is readily accomplished by entering the amendments to the list presented at the screen 414 with a series of amendment options presented in a conventional fashion with the list. As noted, the list may be personally selected by the user in the manner described, or created as a default by network 10.

Collectively, the Jump command, Index command, Directory command, Guide command, and Path command as described enable the user to quickly and easily ascertain the "location" of either the partitioned application presently displayed or the "location" of a desired partitioned application. "Location," as used in reference to the preferred embodiment means the specific relationships that a particular partitioned application bears to other such applications, and the method for selecting particular partitioned applications from such relationships. The techniques for querying a database of objects, embodied in network 10 is an advance over the prior art, insofar as no foreknowledge of either database structure or query technique or syntax is necessary, the structure and search techniques being made manifest to the user in the course of use of the commands.

RS Application Protocol

RS protocol defines the way the RS supports user application conversation (input and output) and the way RS 400 processes a partitioned application. Partitioned applications are constructed knowing that this protocol will be supported unless modified by the application. The protocol is illus-

US 7,072,849 B1

23

trated FIG. 6. The boxes in FIG. 6 identify processing states that the RS 400 passes through and the arrows indicate the transitions permitted between the various states and are annotated with the reason for the transition.

The various states are: (A) Initialize RS, (B) Process Objects, (C) Interpretively Execute Pre-processors, (D) Wait for Event, (E) Process Event, and (F) Interpretively Execute Function Extension and/or Post-processors.

The transitions between states are: (1a) Logon Page Template Object Identification (PTO-id), (1b) Object Identification, (2) Trigger Program Object identification (PO-id) & return, (3) Page Partition Template (PPT) or Window Stack Processing complete, (4) Event Occurrence, and (5) Trigger PO-id and Return.

Transition (1a) from Initialize RS (A) to Process Objects (B) occurs when an initialization routine passes the object-id of the logon PTO to object interpreter 435, when the service is first invoked. Transition (1b) from Process Event (E) to Process Objects (B) occurs whenever a navigation event causes a new page template object identification (PTO-id) to be passed to object interpreter 435; or when an open window event (verb or function key) occurs passing a window object-id to the object interpreter 435; or a close window event (verb or function key) occurs causing the current top-most window to be closed.

While in the process object state, object interpreter 435 will request any objects that are identified by external references in call segments. Objects are processed by parsing and interpreting the object and its segments according to the specific object architecture. As object interpreter 435 processes objects, it builds a linked list structure called a page processing table (PPT), shown in FIG. 10, to reflect the structure of the page, each page partition, Page Element Objects (PEOs) required, program objects (POs) required and each window object (WO) that could be called. Object interpreter 435 requests all objects required to build a page except objects that could be called as the result of some event, such as a HELP window object.

Transition (2) from Process Objects (B) to Interpretively Execute Pre-processors (C) occurs when the object interpreter 435 determines that a pre-processor is to be triggered. Object processor 436 then passes the object-id of the program object to the TBOL interpreter 438. TBOL interpreter 438 uses the RS virtual machine to interpretively execute the program object. The PO can represent either a selector or an initializer. When execution is complete, a transition automatically occurs back to Process Objects (B).

Selectors are used to dynamically link and load other objects such as PEOs or other PDOs based upon parameters that they are passed when they are called. Such parameters are specified in call segments or selector segments. This feature enables RS 400 to conditionally deliver information to the user base upon predetermined parameters, such as his personal demographics or locale. For example, the parameters specified may be the transaction codes required to retrieve the user's age, sex, and personal interest codes from records contained in user profiles stored at the switch/file server layer 200.

Initializers are used to set up the application processing environment for a partitioned application and determine what events RS 400 may respond to and what the action will be.

Transition (3) from Process Objects (B) to Wait for Event (D) occurs when object interpreter 435 is finished processing objects associated with the page currently being built or opening or closing a window on a page. In the Wait for Event state (D), an input manager, which in the preferred form

24

shown includes keyboard manager 434 seen in FIG. 8, accepts user inputs. All keystrokes are mapped from their physical codes to logical keystrokes by the Keyboard Manager 434, representing keystrokes recognized by the RS virtual machine.

When the cursor is located in a field of a page element, keystrokes are mapped to the field and the partitioned external variable (PEV) specified in the page element object (PEO) field definition segment by the cooperative action of keyboard manager, 434 and display manager 461. Certain inputs, such as RETURN or mouse clicks in particular fields, are mapped to logical events by keyboard manager 434, which are called completion (or commit) events. Completion events signify the completion of some selection or specification process associated with the partitioned application and trigger a partition level and/or page level post-processor to process the "action" parameters associated with the user's selection and commit event.

Such parameters are associated with each possible choice 20 or input, and are set up by the earlier interpretive execution of an initializer pre-processor in state (C). Parameters usually specify actions to perform a calculation such as the balance due on an order of several items with various prices using sales tax for the user's location, navigate to PTO-id, 25 open window WO-id or close window. Actions parameters that involve the specification of a page or window object will result in transition (1b) to the Process Objects (B) state after the post-processor is invoked as explained below.

Function keys are used to specify one or more functions 30 which are called when the user strikes these keys. Function keys can include the occurrence of logical events, as explained above. Additionally, certain functions may be "filtered", that is, extended or altered by SET_FUNCTION or TRIGGER_FUNCTION verbs recognized by the RS virtual machine. Function keys cause the PO specified as a parameter of the verb to be interpretively executed whenever that function is called. Applications use this technique to modify or extend the functions provided by the RS.

Transition (5) from Process Event (E) to Interpretively 40 Execute Pre-processors (F) occurs when Process Event State determines that a post-processor or function extension PDO is to be triggered. The id of the program object is then passed to the TBOL interpreter 438. The TBOL interpreter 438 uses the RS virtual machine to interpretively execute the PO. 45 When execution is complete a transition automatically occurs back to Process Event (E).

Reception System Software

The reception system 400 software is the interface 50 between the user of personal computer 405 and interactive network 10. The object of reception system software is to minimize mainframe processing, minimize transmission across the network, and support application extensibility and portability.

RS 400 software is composed of several layers, as shown 55 in FIG. 7. It includes external software 451, which is composed of elements well known to the art such as device drivers, the native operating systems; i.e., MS-DOS, machine-specific assembler functions (in the preferred embodiment; e.g., CRC error checking), and "C" runtime library functions; native software 420; and partitioned applications 410.

Again with reference to FIG. 7, native software 420 is 60 compiled from the "C" language into a target machine-specific executable, and is composed of two components: the service software 430 and the operating environment 450.

US 7,072,849 B1

25

Operating environment **450** is comprised of the Logical Operating System **432**, or LOS; and a multitasker **433**. Service software **430** provides functions specific to providing interaction between the user and interactive network **10**, while the operating environment **450** provides pseudo multitasking and access to local physical resources in support of service software **430**. Both layers of native software **420** contain kernel, or device independent functions **430** and **432**, and machine-specific or device dependent functions **433**. All device dependencies are in code resident at RS **400**, and are limited to implementing only those functions that are not common across machine types, to enable interactive network **10** to provide a single data stream to all makes of personal computer which are of the IBM or IBM compatible type. Source code for the native software **420** is included in parent application Ser. No. 388,156 now issued as U.S. Pat. No. 5,347,632, the contents of which patent are incorporated herein by reference. Those interested in a more detailed description of the reception system software may refer to the source code provided in the referenced patent.

Service software **430** is comprised of modules, which are device-independent software components that together obtain, interpret and store partitioned applications existing as a collection of objects. The functions performed by, and the relationship between, the service software **430** module is shown in FIG. 8 and discussed further below.

Through facilities provided by LOS **432** and multitasker **433**, here called collectively operating environment **450**, device-independent multitasking and access to local machine resources, such as multitasking, timers, buffer management, dynamic memory management, file storage and access, keyboard and mouse input, and printer output are provided. The operating environment **450** manages communication and synchronization of service software **430**, by supporting a request/response protocol and managing the interface between the native software **420** and external software **437**.

Applications software layer **410** consists of programs and data written in an interpretive language, "TRINTEX Basic Object Language" or "TBOL," described above. TBOL was written specifically for use in RS **400** and interactive network **10** to facilitate videotext-specific commands and achieve machine-independent compiling. TBOL is constructed as objects, which in interaction with one another comprise partitioned applications.

RS native software **420** provides a virtual machine interface for partitioned applications, such that all objects comprising partitioned applications "see" the same machine. RS native software provides support for the following functions: (1) keyboard and mouse input; (2) text and graphics display; (3) application interpretation; (4) application database management; (5) local application storage; (6) network and link level communications; (7) user activity data collection; and (8) advertisement management.

With reference to FIG. 8, service software **430** is comprised of the following modules: start-up (not shown); keyboard manger **434**; object interpreter **435**; TBOL interpreter **438**; object storage facility **439**; display manager **461**; data collection manager **441**; ad manager **442**; object/communications manager interface **443**; link communications manager **444**; and fatal error manager **469**. Each of these modules has responsibility for managing a different aspect of RS **400**.

Startup reads RS **400** customization options into RAM, including modem, device driver and telephone number options, from the file CONFIG.SM. Startup invokes all RS **400** component startup functions, including navigation to

26

the first page, a logon screen display containing fields initialized to accept the user's id and password. Since Startup is invoked only at initialization, for simplicity, it has not been shown in FIG. 8.

The principal function of keyboard manger **434** is to translate personal computer dependent physical input into a consistent set of logical keys and to invoke processors associated with these keys. Depending on the LOS key, and the associated function attached to it, navigation, opening of windows, and initiation of filter or post-processor TBOL programs may occur as the result input events handled by the keyboard manger **434**. In addition, keyboard manger **434** determines inter and intra field cursor movement, and coordinates the display of field text and cursor entered by the user with display manager **461**, and sends information regarding such inputs to data collection manager **441**.

Object interpreter **435** is responsible for building and recursively processing a table called the "Page Processing Table," or PPT. Object interpreter **435** also manages the opening and closing of windows at the current page. Object interpreter **435** is implemented as two sub-components: the object processor **436** and object scanner **437**.

Object processor **436** provides an interface to keyboard manger **434** for navigation to new pages, and for opening and closing windows in the current page. Object processor **436** makes a request to object storage facility **439** for a page template object (PTO) or window object (WO), as requested by keyboard manger **434**, and for objects and their segments which comprise the PTO or WO returned by object storage facility **439** to object processor **436**. Based on the particular segments comprising the object(s) making up the new PTO or WO, object processor **436** builds or adds to the page processing table (PPT), which is an internal, linked-list, global data structure reflecting the structure of the page or page format object (PFO), each page partition or page element object (PEO), and program objects (POs) required and each window object (WO) that could be called. Objects are processed by parsing and interpreting each object and its-segment(s) according to their particular structure as formalized in the data object architecture (DOA). While in the process object state, (state "B" of FIG. 6), object processor **436** will request any objects specified by the PTO that are identified by external references in call segments (e.g. field

level program call **518**, page element selector call **524**, page format call **526** program call **532**, page element call **522** segments) of such objects, and will, through a request to TBOL interpreter **438**, fire initializers and selectors contained in program data segments of all PTO constituent program objects, at the page, element, and field levels. Object processor **436** requests all objects required to build a page, except objects that could only be called as the result of some event external to the current partitioned application, such as a HELP window object. When in the course of

building or adding to the PPT and opening/closing WOs, object processor encounters a call to an "ADSLOT" object id, the next advertisement object id at ad manager **442** is fetched, and the identified advertisement object is retrieved either locally, if available, or otherwise from the network, so that the presentation data for the advertisement can be sent to display manager **461** along with the rest of the presentation data for the other objects to enable display to the user. Object processor **436** also passes to data collection manager **441** all object ids that were requested and object ids that were viewed. Upon completion of page or window processing, object processor **436** enters the wait for event state, and control is returned to keyboard manger **434**.

US 7,072,849 B1

27

The second component of object interpreter **435**, object scanner **437**, provides a file-like interface, shared with object storage facility **439**, to objects currently in use at RS **400**, to enable object processor **436** to maintain and update the PPT. Through facilities provided by object scanner **437**, object processor recursively constructs a page or window in the requested or current partitioned application, respectively.

Object storage facility **439** provides an interface through which object interpreter **435** and TBOL interpreter **438** either synchronously request (using the TBOL verb operator "GET") objects without which processing in either module cannot continue, or asynchronously request (using the TBOL verb operator "FETCH") objects in anticipation of later use. Object storage facility **439** returns the requested objects to the requesting module once retrieved from either local store **440** or interactive network **10**. Through control structures shared with the object scanner **437**, object storage facility determines whether the requested object resides locally, and if not, makes an attempt to obtain it from interactive network **10** through interaction with link communications manager **444** via object/communications manager interface **443**.

When objects are requested from object storage facility **439**, only the latest version of the object will be provided to guarantee currency of information to the user. Object storage facility **439** assures currency by requesting version verification from network **10** for those objects which are available locally and by requesting objects which are not locally available from delivery system **20** where currency is maintained.

Version verification increases response time. Therefore, not all objects locally available are version checked each time they are requested. Typically, objects are checked only the first time they are requested during a user session. However, there are occasions, as for example in the case of objects relating to news applications, where currency is always checked to assure integrity of the information.

The frequency with which the currency of objects is checked depends on factors such as the frequency of updating of the objects. For example, objects that are designated as ultrastable in a storage control parameter in the header of the object are never version checked unless a special version control object sent to the RS as part of logon indicates that all such objects must be version checked. Object storage facility **439** marks all object entries with such a stability category in all directories indicating that they must be version checked the next time they are requested.

Object storage facility **439** manages objects locally in local store **440**, comprised of a cache (segmented between available RAM and a fixed size disk file), and stage (fixed size disk file). Ram and disk cached objects are retained only during user sessions, while objects stored in the stage file are retained between sessions. The storage control field, located in the header portion of an object, described more fully hereafter as the object "storage candidacy", indicates whether the object is stageable, cacheable or trashable.

Stageable objects must not be subject to frequent change or update. They are retained between user sessions on the system, provided storage space is available and the object has not discarded by a least-recently-used (LRU) algorithm of a conventional type; e.g., see *Operating System Theory*, by Coffman, Jr. and Denning, Prentice Hall Publishers, New York, 1973, which in accordance with the design of network **10**, operates in combination with the storage candidacy value to determine the object storage priority, thus rendering the stage self-configuring as described more fully hereafter. Over time, the self-configuring stage will have the effect of retaining within local disk storage those objects which the

28

user has accessed most often. The objects retained locally are thus optimized to each individual user's usage of the applications in the system. Response time to such objects is optimized since they need not be retrieved from the interactive computer system.

Cacheable objects can be retained during the current user session, but cannot be retained between sessions. These objects usually have a moderate update frequency. Object storage facility **439** retains objects in the cache according to the LRU storage retention algorithm. Object storage facility **439** uses the LRU algorithm to ensure that objects that are least frequently used forfeit their storage to objects that are more frequently used.

Trashable objects can be retained only while the user is in the context of the partitioned application in which the object was requested. Trashable objects usually have a very high update frequency and must not be retained to ensure that the user has access to the most current data.

More particularly and, as noted above, in order to render a public informational and transactional network of the type considered here attractive, the network must be both economical to use and fast. That is to say, the network must supply information and transactional support to the user at minimal costs and with a minimal response time. These objectives are sought to be achieved by locating as many information and transactional support objects which the user is likely to request, as close to the user as possible; i.e., primarily at the user's RS **400** and secondarily at delivery system **20**. In this way, the user will be able to access objects required to support a desired application with minimal intervention of delivery system **20**, thus reducing the cost of the session and speeding the response time.

However, the number of objects that can be maintained at RS **400** is restricted by at least two factors: the RS **400** storage capacity; i.e., RAM and disk sizes, and the need to maintain the stored objects current.

In order to optimize the effectiveness of the limited storage space at RS **400**, the collection of objects is restricted to those likely to be requested by the user; i.e., tailored to the user's tastes—and to those least likely to be time sensitive; i.e., objects which are stable. To accomplish this, objects are coded for storage candidacy to identify when they will be permitted at RS **400**, and subject to the LRU algorithm to maintain presence at RS **400**. Additionally, to assure currency of the information and transaction support provided at RS **400**, objects are further coded for version identification and checking in accordance with a system of priorities that are reflected in the storage candidacy coding.

Specifically, to effect object storage management, objects are provided with a coded version id made up of the storage control byte and version control bytes identified above as elements of the object header, specifically, bytes **16** and **18** shown in FIG. **4b**. In preferred form, the version id is comprised of bytes **16** and **18** to define two fields, a first 13 bit field to identify the object version and a second three bit field to identify the object storage candidacy.

In this arrangement, the storage candidacy value of the object is addressed to not only the question of storage preference but also object currency. Specifically, the storage candidacy value establishes the basis upon which the object will be maintained at RS **400** and also identifies the susceptibility of the object to becoming stale by dictating when the object will be version checked to determine currency.

The version value of the object on the other hand, provides a parameter that can be checked against predetermined values available from delivery system **20** to deter-

US 7,072,849 B1

29

mine whether an object stored at RS **400** is sufficiently current to permit its continued use, or whether the object has become stale and needs to be replaced with a current object from delivery system **20**.

Still further, object storage management procedure further includes use of the LRU algorithm, for combination with the storage and version coding to enable discarding of objects which are not sufficiently used to warrant retention, thus personalizing the store of objects at RS **400** to the user's tastes. Particularly, object storage facility **439**, in accordance with the LRU algorithm maintains a usage list for objects. As objects are called to support the user's applications requests, the objects are moved to the top of a usage list. As other objects are called, they push previously called objects down in the list. If an object is pushed to the bottom of the list before being recalled, it will be forfeited from the list if necessary to make room for the next called object. As will be appreciated, should a previously called object be again called before it is displaced from the list, it will be promoted to the top of the list, and once more be subject to depression in the list and possible forfeiture as other objects are called.

As pointed out above, in the course of building the screens presented to the user, objects will reside at various locations in RS **400**. For example, objects may reside in the RS **400** RAM where the object is supporting a particular application screen then running or in a cache maintained at either RAM or disk **424** where the object is being held for an executing application or staged on the fixed size file on disk **424** noted above where the object is being held for use in application likely to be called by the user in the future.

In operation, the LRU algorithm is applied to all these regions and serves to move an object from RAM cache to disk cache to disk file, and potentially off RS **400** depending on object usage.

With regard to the storage candidacy value, in this arrangement, the objects stored at RS **400** include a limited set of permanent objects; e.g., those supporting logon and logoff, and other non-permanent objects which are subject to the LRU algorithm to determine whether the objects should be forfeited from RS **400** as other objects are added. Thus, in time, and based on the operation of the LRU algorithm and the storage candidacy value, the collection of objects at RS **400** will be tailored to the usage characteristics of the subscriber; i.e., self-configuring.

More particularly, the 3-bit field of the version id that contains the storage candidacy parameter can have 8 different values. A first candidacy value is applied where the object is very sensitive to time; e.g., news items, volatile pricing information such as might apply to stock quotes, etc. In accordance with this first value, the object will not be permitted to be stored on RS **400**, and RS **400** will have to request such objects from delivery system **20** each time it is accessed, thus, assuring currency. A second value is applied where the object is sensitive to time but less so than the first case; e.g., the price of apples in a grocery shopping application. Here, while the price might change from day to day, it is unlikely to change during a session. Accordingly the object will be permitted to persist in RAM or at the disk cache during a session, but will not be permitted to be maintained at RS **400** between sessions.

Continuing down the hierarchy of time sensitivity, where the object concerns information sufficiently stable to be maintained between sessions, a third storage candidacy value is set to permit the object to be stored at RS **400** between sessions, on condition that the object will be version check the first time it is accessed in a subsequent session. As will be appreciated, during a session, and under

30

the effect of the LRU algorithm, lack of use at RS **400** of the object may result in it being forfeited entirely to accommodate new objects called for execution at RS **400**.

Still further, a fourth value of storage candidacy is applied where the object is considered sufficiently stable as not to require version checking between sessions; e.g., objects concerning page layouts not anticipated to change. In this case, the storage candidacy value may be encoded to permit the object to be retained from session to session without version checking. Here again, however, the LRU algorithm may cause the object to forfeit its storage for lack of use.

Where the object is of a type required to be stored at RS **400**, as for example, objects needed to support standard screens, it is coded for storage between sessions and not subject to the LRU algorithm forfeiture. However, where such objects are likely to change in the future they may be required to be version checked the first time they are accessed in a session and thus be given a fifth storage candidacy value. If, on the other hand, the required stored object is considered likely to be stable and not require even version checking; e.g., logon screens, it will be coded with a sixth storage candidacy value for storage without version checking so as to create a substantially permanent object.

Continuing, where a RS **400** includes a large amount of combined RAM and disk capacity, it would permit more objects to be stored. However, if objects were simply coded in anticipation of the larger capacity, the objects would potentially experience difficulty, as for example, undesired forfeiture due to capacity limitations if such objects were supplied to RS **400** units having smaller RAM and disk sizes. Accordingly, to take advantage of the increased capacity of certain RS **400** units without creating difficulty in lower capacity units, objects suitable for storage in large capacity units can be so coded for retention between sessions with a seventh and eighth storage candidacy value depending upon whether the stored large capacity object requires version checking or not. Here, however, the coding will be interpreted by smaller capacity units to permit only cacheable storage to avoid undesirable forfeiture that might result from over filling the smaller capacity units.

Where an object is coded for no version checking need may nonetheless arise for a version check at some point. To permit version checking of such objects, a control object is provided at RS **400** that may be version checked on receipt of a special communication from delivery system **20**. If the control object fails version check, then a one shot version checking attribute is associated with all existing objects in RS **400** that have no version checking attributes. Thereafter, the respective objects are version checked, the one shot check attribute is removed and the object is caused to either revert to its previous state if considered current or be replaced if stale.

Still further, objects required to be stored at RS **400** which are not version checked either because of lack of requirement or because of no version check without a control object, as described above, can accumulate in RS **400** as dead objects. To eliminate such accumulation, all object having required storage are version checked over time. Particularly, the least recently used required object is version checked during a session thus promoting the object to the top of the usage list if it is still to be retained at RS **400**. Accordingly, one such object will be checked per session and over time, all required objects will be version checked thereby eliminating the accumulation of dead objects.

However, in order to work efficiently, the version check attribute of the object should be ignored, so that even required object can be version checked. Yet, in certain

US 7,072,849 B1

31

circumstances, e.g., during deployment of new versions of the reception system software containing new objects not yet supported on delivery system **20** which may be transferred to the fixed storage file of RS **400** when the new version is loaded, unconditional version checking may prematurely deletes the object from the RS **400** as not found on delivery system **20**. To avoid this problem, a sweeper control segment in the control object noted above can be used to act as a switch to turn the sweep of dead objects on and off.

With respect to version checking for currency, where an object stored at RS **400** is initially fetched or accessed during a session, a request to delivery system **20** is made for the object by specifying the version id of the object stored at RS **400**.

In response, delivery system **20** will advise the reception system **400** either that the version id of the stored object matches the currency value; i.e., the stored object is acceptable, or deliver a current object that will replace the stored object shown to be stale. Alternatively, the response may be that the object was not found. If the version of the stored object is current, the stored object will be used until verified again in accordance with its storage candidacy. If the stored object is stale, the new object delivered will replace the old one and support the desired screen. If the response is object not found, the stored object will be deleted.

Therefore, based on the above description, network **10** is seen to include steps for execution at storage facility **439** which enables object reception, update and deletion by means of a combination of operation of the LRU algorithm and interpretation of the storage candidacy and version control values. In turn, these procedures cooperate to assure a competent supply of objects at RS **400** so as to reduce the need for intervention of delivery system **20**, thus reducing cost of information supply and transactional support so as to speed the response to user requests.

TBOL interpreter **438** shown in FIG. 8 provides the means for executing program objects, which have been written using an interpretive language, TBOL described above. TBOL interpreter **438** interprets operators and operand contained in program object **508**, manages TBOL variables and data, maintains buffer and stack facilities, and provides a runtime library of TBOL verbs.

TBOL verbs provide support for data processing, program flow control, file management, object management, communications, text display, command bar control, open/close window, page navigation and sound. TBOL interpreter also interacts with other native modules through commands contained in TBOL verbs. For example: the verb "navigate" will cause TBOL interpreter **438** to request object interpreter **435** to build a PPT based on the PTO id contained in the operand of the NAVIGATE verb; "fetch" or "GET" will cause TBOL interpreter **438** to request an object from object storage facility **439**; "SET_FUNCTION" will assign a filter to events occurring at the keyboard manger **434**; and "FORMAT," "SEND," and "RECEIVE" will cause TBOL interpreter **438** to send application level requests to object/communications manager interface **433**.

Data areas managed by TBOL interpreter **438** and available to TBOL programs are Global External Variables (GEVs), Partition External Variables (PEVs), and Runtime Data Arrays (RDAs).

GEVs contain global and system data, and are accessible to all program objects as they are executed. GEVs provide a means by which program objects may communicate with other program objects or with the RS native code, if declared in the program object. GEVs are character string variables that take the size of the variables they contain. GEVs may

32

preferably contain a maximum of 32,000 variables and are typically used to store such information as program return code, system date and time, or user sex or age. TBOL interpreter **438** stores such information in GEVs when requested by the program which initiated a transaction to obtain these records from the RS or user's profile stored in the interactive system.

Partition external variables (PEVs) have a scope restricted to the page partition on which they are defined. PEVs are used to hold screen field data such that when PEOs and window objects are defined, the fields in the page partitions with which these objects are to be associated are each assigned to a PEV. When applications are executed, TBOL interpreter **438** transfers data between screen fields and their associated PEV. When the contents of a PEV are modified by user action or by program direction, TBOL interpreter **428** makes a request to display manager **461** to update the screen field to reflect the change. PEVs are also used to hold partition specific application data, such as tables of information needed by a program to process an expected screen input.

Because the scope of PEVs is restricted to program objects associated with the page partition in which they are defined, data that is to be shared between page partitions or is to be available to a page-level processor must be placed in GEVs or RDAs.

RDAs are internal stack and save buffers used as general program work areas. RDAs are dynamically defined at program object "runtime" and are used for communication and transfer of data between programs when the data to be passed is not amenable to the other techniques available. Both GEVs and RDAs include, in the preferred embodiment, 8 integer registers and 8 decimal registers. Preferably, there are also 9 parameter registers limited in scope to the current procedure of a program object.

All variables may be specified as operand of verbs used by the virtual machine. The integer and decimal registers may be specified as operand for traditional data processing. The parameter registers are used for passing parameters to "called" procedures. The contents of these registers are saved on an internal program stack when a procedure is called, and are restored when control returns to the "calling" procedure from the "called" procedure.

TBOL interpreter **438**, keyboard manger **434**, object interpreter **435**, and object storage facility **439**, together with device control provided by operating environment **450**, have principal responsibility for the management and execution of partitioned applications at the RS **400**. The remaining native code modules function in support and ancillary roles to provide RS **400** with the ability display partitioned applications to the user (display manager **461**), display advertisements (ad manager **442**), to collect usage data for distribution to interactive network **10** for purposes of targeting such advertisements (data collection manager **441**), and prepare for sending, and send, objects and messages to interactive network **10** (object/communications manager interface **443** and link communications manager **444**). Finally, the fatal error manager exists for one purpose: to inform the user of RS **400** and transmit to interactive network **10** the inability of RS **400** to recover from a system error.

Display manager **461** interfaces with a decoder using the North American Presentation Level Protocol Syntax (NAPLPS), a standard for encoding graphics data, or text code, such as ASCII, which are displayed on monitor **412** of the user's personal computer **405** as pictorial codes. Codes for other presentation media, such as audio, can be specified by

US 7,072,849 B1

33

using the appropriate type code in the presentation data segments. Display manager 461 supports the following functions: send NAPLPS strings to the decoder; echo text from a PEV; move the cursor within and between fields; destructive or non-destructive input field character deletion; "ghost" and "unghost" fields (a ghosted field is considered unavailable, unghosted available); turn off or on the current field cursor; open, close, save and restore bit maps for a graphics window; update all current screen fields by displaying the contents of their PEVs, reset the NAPLPS decoder to a known state; and erase an area of the screen by generating and sending NAPLPS to draw a rectangle over that area. Display manager 461 also provides a function to generate a beep through an interface with a machine-dependent sound driver.

In accordance with the method of the present invention, Ad manager 442 is invoked by object interpreter 435 to return the object id of the next available advertisement to be displayed. Ad manager 442 maintains a queue of advertising object id's targeted to the specific user currently accessing interactive network 10. Advertising objects are pre-fetched from interactive system 10 from a personalized queue of advertising ids that is constructed using data previously collected from user generated events and/or reports of objects used in the building of pages or windows, compiled by data collection manager 466 and transmitted to interactive system 10.

Advertising objects 510 are PEOs that, through user invocation of a "LOOK" command, cause navigation to partitioned applications that may themselves support, for example, ordering and purchasing of merchandise.

An advertising object id list, or "ad queue," is requested in a transaction message to delivery system 20 by ad manager 442 immediately after the initial logon response. The logon application at RS 400 places the advertising list in a specific RS global storage area called a SYS_GEV (system global external variable), which is accessible to all applications as well as to the native RS code). The Logon application also obtains the first two ad object id's from the queue and provides them to object storage facility 439 so the advertising objects can be requested. However, at logon, since no advertising objects are available at RS local storage facilities 440, ad objects, in accordance with the described storage candidacy, not being retained at the reception system between sessions, they must be requested from interactive network 10.

In a preferred embodiment, the following parametric values are established for ad manager 442: advertising object is queue capacity, replenishment threshold for advertising object id's and replenishment threshold for number of outstanding pre-fetched advertising objects. These parameters are set up in GEVs of the RS virtual machine by the logon application program object from the logon response from high function system 110. The parameters are then also accessible to the ad manager 442. Preferred values are an advertising queue capacity of 15, replenishment value of 10 empty queue positions and a pre-fetched advertising object threshold of 3.

Ad manager 442 pre-fetches advertising objects by passing advertising object id's from the advertising queue to object storage facility 439 which then retrieves the object from the interactive system if the object is not available locally. Advertising objects are pre-fetched, so they are available in RS local store 440 when requested by object interpreter 435 as it builds a page. The ad manager 442 pre-fetches additional advertising objects whenever the number of pre-fetched advertising objects not called by

34

object interpreter 435; i.e. the number of remaining advertising objects, falls below the pre-fetch advertising threshold.

Whenever the advertising object id queue has more empty positions than replenishment threshold value, a call is made to the advertising object id queue application in high function system 110 shown in FIG. 2, via object/communications manager interface 443 for a number of advertising object id's equal to the threshold value. The response message from system 110 includes a list of advertising object id's, which ad manager 442 enqueues.

Object interpreter 435 requests the object id of the next advertising object from ad manager 442 when object interpreter 435 is building a page and encounters an object call for a partition and the specified object-id equals the code word, "ADSLOT." If this is the first request for an advertising object id that ad manager 442 has received during this user's session, ad manager 442 moves the advertising object id list from the GEV into its own storage area, which it uses as an advertising queue and sets up its queue management pointers, knowing that the first two advertising objects have been pre-fetched.

Ad manager 442 then queries object storage facility 439, irrespective of whether it was the first request of the session. The query asks if the specified advertising object id pre-fetch has been completed, i.e., is the object available locally at the RS. If the object is available locally, the object-id is passed to object interpreter 435, which requests it from object storage facility 439. If the advertising object is not available in local store 440, ad manager 442 attempts to recover by asking about the next ad that was pre-fetched. This is accomplished by swapping the top and second entry in the advertising queue and making a query to object storage facility 439 about the new top advertising object id. If that object is not yet available, the top position is swapped with the third position and a query is made about the new top position.

Besides its ability to provide advertising that have been targeted to each individual user, two very important response time problems have been solved by ad manager 442 of the present invention. The first is to eliminate from the new page response time the time it takes to retrieve an advertising object from the host system. This is accomplished by using the aforementioned pre-fetching mechanism.

The second problem is caused by pre-fetching, which results in asynchronous concurrent activities involving the retrieval of objects from interactive system 10. If an advertising object is pre-fetched at the same time as other objects required for a page are requested, the transmission of the advertising object packets could delay the transmission of the other objects required to complete the current page by the amount of time required to transmit the advertising object(s). This problem is solved by the structuring the requests from object interpreter 435 to the ad manager 442 in the following way:

1. Return next object id of pre-fetched advertising object & pre-fetch another;
2. Return next advertising object id only; and
3. Pre-fetch next advertising object only.

By separating the function request (1) into its two components, (2) and (3), object interpreter 435 is now able to determine when to request advertising object id's and from its knowledge of the page build process, is able to best determine when another advertising object can be pre-fetched, thus causing the least impact on the page response time. For example, by examining the PPT, object interpreter 435 may determine whether any object requests are out-

US 7,072,849 B1

35

standing. If there are outstanding requests, advertising request type 2 would be used. When all requested objects are retrieved, object interpreter 435 then issues an advertising request type 3. Alternatively, if there are no outstanding requests, object interpreter 435 issues an advertising request type 1. This typically corresponds to the user's "think time" while examining the information presented and when RS 400 is in the Wait for Event state (D).

Data collection manager 441 is invoked by object interpreter 435 and keyboard manager 434 to keep records about what objects a user has obtained (and, if a presentation data segment 530 is present, seen) and what actions users have taken (e.g. "NEXT," "BACK," "LOOK," etc.)

The data collection events that are to be reported during the user's session are sensitized during the logon process. The logon response message carries a data collection indicator with bit flags set to "on" for the events to be reported. These bit flags are enabled (on) or disabled (off) for each user based on information contained in the user's profile stored and sent from high function host 110. A user's data collection indicator is valid for the duration of his session. The type of events to be reported can be changed at will in the host data collection application. However, such changes will affect only users who logon after the change.

Data collection manager 441 gathers information concerning a user's individual system usage characteristics. The types of informational services accessed, transactions processed, time information between various events, and the like are collected by data collection manager 441, which compiles the information into message packets (not shown). The message packets are sent to network 10 via object/communication manager interface 443 and link communications manager 444. Message packets are then stored by high function host 110 and sent to an offline processing facility for processing. The characteristics of users are ultimately used as a means to select or target various display objects, such as advertising objects, to be sent to particular users based on consumer marketing strategies, or the like, and for system optimization.

Object/communications manager interface 443 is responsible for sending and receiving DIA (Data Interchange Architecture described above) formatted messages to or from interactive network 10. Object/communications manager 443 also handles the receipt of objects, builds a DIA header for messages being sent and removes the header from received DIA messages or objects, correlates requests and responses, and guarantees proper block sequencing. Object/communications manager interface 443 interacts with other native code modules as follows: object/communications manager 443 (1) receives all RS 400 object requests from object storage facility 439, and forwards objects received from network 10 via link communications manager 444 directly to the requesting modules; (2) receives ad list requests from ad manager 442, which thereafter periodically calls object/communications manager 443 to receive ad list responses; (3) receives data collection messages and send requests from data collection manager 441; (4) receives application-level requests from TBOL interpreter 438, which also periodically calls object/communications manager interface 443 to receive responses (if required); and (5) receives and sends DIA formatted objects and messages from and to link communications manager 444.

Object/communications manager interface 443 sends and receives DIA formatted messages on behalf of TBOL interpreter 438 and sends object requests and receives objects on behalf of object storage facility 439. Communication packets received containing parts of requested objects are passed

36

to object storage facility 439 which assembles the packets into the object before storing it. If the object was requested by object interpreter 435, all packets received by object storage facility 439 are also passed to object interpreter 435 avoiding the delay required to receive an entire object before processing the object. Objects which are pre-fetched are stored by object storage facility 439.

Messages sent to interactive network 10 are directed via DIA to applications in network 10. Messages may include transaction requests for records or additional processing of records or may include records from a partitioned application program object or data collection manager 441. Messages to be received from network 10 usually comprise records requested in a previous message sent to network 10. Requests received from object storage facility 439 include requests for objects from storage in interactive system 10. Responses to object requests contain either the requested object or an error code indicating an error condition.

Object/communications manager 443 is normally the exclusive native code module to interface with link communications manager 444 (except in the rare instance of a fatal error). Link communications manager 444 controls the connecting and disconnecting of the telephone line, telephone dialing, and communications link data protocol. Link communications manager 444 accesses network 10 by means of a communications medium (not shown) link communications manager 444, which is responsible for a dial-up link on the public switched telephone network (PSTN). Alternatively, other communications means, such as cable television or broadcast media, may be used. Link communications manager 444 interfaces with TBOL interpreter for connect and disconnect, and with interactive network 10 for send and receive.

Link communications manager 444 is subdivided into modem control and protocol handler units. Modem control (a software function well known to the art) hands the modem specific handshaking that occurs during connect and disconnect. Protocol handler is responsible for transmission and receipt of data packets using the TCS (TRINTEX Communications Subsystem) protocol (which is a variety of OSI link level protocol, also well known to the art).

Fatal error manager 469 is invoked by all reception system components upon the occurrence of any condition which precludes recovery. Fatal error manager 469 displays a screen to the user with a textual message and an error code through display manager 461. Fatal error manager 469 sends an error report message through the link communications manager 444 to a subsystem of interactive network 10.

The source code for the reception system software as noted above is described in parent application Ser. No. 388,156 filed Jul. 28, 1989, now issued as U.S. Pat. No. 5,347,632, the contents of which are incorporated herein by reference.

Sample Application

Page 255 illustrated in FIG. 3b corresponds to a partitioned application that permits a user to purchase apples. It shows how the monitor screen 414 of the reception system 400 might appear to the user. Displayed page 255 includes a number of page partitions and corresponding page elements.

The page template object (PTO) 500 representing page 255 is illustrated in FIG. 9. PTO 500 defines the composition of the page, including header 250, body 260, display fields 270, 271, 272, advertising 280, and command bar 290. Page element objects (PEOs) 504 are associated with page parti-

US 7,072,849 B1

37

tions numbered; e.g., 250, 260, 280. They respectively, present information in the header 250, identifying the page topic as ABC APPLES; in the body 260, identifying the cost of apples; and prompt the user to input into fields within body 260 the desired number of apples to be ordered. In advertising 280, presentation data and a field representing a post-processor that will cause the user to navigate to a targetable advertising, is presented.

In FIG. 9, the structure of PTO 500 can be traced. PTO 500 contains a page format call segment 526, which calls page format object (PFO) 502. PFO 502 describes the location and size of partitions on the page and numbers assigned to each partition. The partition number is used in page element call segments 522 so that an association is established between a called page element object (PEO) 504 and the page partition where it is to be displayed. Programs attached to this PEO can be executed only when the cursor is in the page partition designated within the PEO.

PTO 500 contains two page element call segments 522, which reference the PEOs 504 for partitions 250 and 260. Each PEO 504 defines the contents of the partition. The header in partition 250 has only a presentation data segment 530 in its PEO 504. No input, action, or display fields are associated with that partition.

The PEO 504 for partition 260 contains a presentation data segment 530 and field definition segments 516 for the three fields that are defined in that partition. Two of the fields will be used for display only. One field will be used for input of user supplied data.

In the example application, the PEO 504 for body partition 260 specifies that two program objects 508 are part of the body partition. The first program, shown in Display field 270, 271, 272, is called an initializer and is invoked unconditionally by TBOL interpreter 438 concurrently with the display of presentation data for the partition. In this application, the function of the initializer is represented by the following pseudo-code:

1. Move default values to input and display fields;
2. "SEND" a transaction to the apple application that is resident on interactive system 10;
3. "RECEIVE" the result from interactive system 10; i.e. the current price of an apple;
4. Move the price of an apple to PEV 271 so that it will be displayed;
5. Position the cursor on the input field; and
6. Terminate execution of this logic.

The second program object 508 is a field post-processor. It will be invoked conditionally, depending upon the user keystroke input. In this example, it will be invoked if the user changes the input field contents by entering a number. The pseudo code for this post-processor is as follows:

1. Use the value in PEV 270 (the value associated with the data entered by the user into the second input data field 270) to be the number of apples ordered.
2. Multiply the number of apples ordered times the cost per apple previously obtained by the initializer;
3. Construct a string that contains the message "THE COST OF THE APPLES YOU ORDERED IS \$45.34;"
4. Move the string into PEV 272 so that the result will be displayed for the user; and
5. Terminate execution of this logic.

The process by which the "APPLES" application is displayed, initialized, and run is as follows.

The "APPLES" application is initiated when the user navigates from the previous partitioned application, with the navigation target being the object id of the "APPLES" PTO 500 (that is, object id ABC1). This event causes keyboard

38

manager 434 to pass the PTO object id, ABC1 (which may, for example, have been called by the keyword navigation segment 520 within a PEO 504 of the previous partitioned application), to object interpreter 435. With reference to the RS application protocol depicted in FIG. 6, when the partitioned application is initiated, RS 400 enters the Process Object state (B) using transition (1). Object interpreter 435 then sends a synchronous request for the PTO 500 specified in the navigation event to object storage facility 439. Object storage facility 439 attempts to acquire the requested object from local store 440 or from delivery system 20 by means of object/communication manager 443, and returns an error code if the object cannot be acquired.

Once the PTO 500 is acquired by object/communications manager 443, object interpreter 435 begins to build PPT by parsing PTO 500 into its constituent segment calls to pages and page elements, as shown in FIG. 4d and interpreting such segments. PFO and PEO call segments 526 and 522 require the acquisition of the corresponding objects with object id's <ABC>, <ABCX> and <ABCY>. Parsing and interpretation of object ABCY requires the further acquisition of program objects <ABC1> and <ABCJ>.

During the interpretation of the PEOs 504 for partitions 250 and 260, other RS 400 events are triggered. This corresponds to transition (2) to interpret pre-processors state (C) in FIG. 6. Presentation data 530 is sent to display manager 461 for display using a NAPLPS decoder within display manager 461, and, as the PEO <ABCY> for partition 260 is parsed and interpreted by object interpreter 435, parameters in program call segment 532 identify the program object <ABC1> as an initializer. Object interpreter 435 obtains the program object from object storage facility 439, and makes a request to TBOL interpreter 438 to execute the initializer program object 508 <ABC1>. The initializer performs the operations specified above using facilities of the RS virtual machine. TBOL interpreter 438, using operating environment 450, executes initializer program object 506 <ABC1>, and may, if a further program object 508 is required in the execution of the initializer, make a synchronous application level object request to object storage facility 439. When the initializer terminates, control is returned to object interpreter 435, shown as the return path in transition (2) in FIG. 6.

Having returned to the process object state (B), object processor 435 continues processing the objects associated with PTO <ABC1>. Object interpreter continues to construct the PPT, providing RS 400 with an environment for subsequent processing of the PTO <ABC1> by pre-processors and post-processors at the page, partition, and field levels. When the PPT has been constructed and the initializer executed, control is returned to keyboard manager 434, and the RS enters the wait for event (E) State, via transition (4), as shown in FIG. 6.

In the wait for event state, the partitioned application waits for the user to create an event. In any partitioned application, the user has many options. For example, the user may move the cursor to the "JUMP" field 296 on the command bar 290, which is outside the current application, and thus cause subsequent navigation to another application.

For purposes of this example, it is assumed that the user enters the number of apples he wishes to order by entering a digit in display field 271.

Keyboard manager 434 translates the input from the user's keyboard to a logical representation independent of any type of personal computer. Keyboard manager 434 saves the data entered by the user in a buffer associated with the current field defined by the location of the cursor. The

US 7,072,849 B1

39

buffer is indexed by its PEV number, which is the same as the field number assigned to it during the formation of the page element. Keyboard manager 434 determines for each keystroke whether the keystroke corresponds to an input event or to an action or completion event. Input events are logical keystrokes and are sent by keyboard manager to display manager 461, which displays the data at the input field location. Display manager 461 also has access to the field buffer as indexed by its PEV number.

The input data are available to TBOL interpreter 438 for subsequent processing. When the cursor is in a partition, only the PEVs for that partition are accessible to the RS virtual machine. After the input from the user is complete (as indicated by a user action such as pressing the RETURN key or entry of data into a field with an action attribute), RS 400 enters the Process Event state (E) via transition (4).

For purposes of this example, let us assume that the user enters the digit “5” in input field 270. A transition is made to the process event state (E). Keyboard manager 434 and display manager 437 perform a number of actions, such as the display of the keystroke on the screen, the collection of the keystroke for input, and optionally, the validation of the keystroke, i.e. numeric input only in numeric fields. When the keystroke is processed, a return is made to the wait for event state (D) Edit attributes are specified in the field definition segment.

Suppose the user inputs a “6” next. A transition occurs to the PE state and after the “6” is processed, the Wait for Event (D) state is reentered. If the user hits the “completion” key (e.g., ENTER) the Process Event (E) state will be entered. The action attributes associated with field 272 identify this as a system event to trigger post-processor program object <ABCJ>. When the interpretive execution of program object <ABCJ> is complete, the wait for event state (D) will again be entered. The user is then free to enter another value in the input field, or select a command bar function and exit the apples application.

While this invention has been described in its preferred form, it will be appreciated that changes may be made in the form, construction, procedure and arrangement of its various elements and steps without departing from its spirit or scope.

We claim:

1. A method for presenting advertising obtained from a computer network, the network including a multiplicity of user reception systems at which respective users can request applications, from the network, that include interactive services, the respective reception systems including a monitor at which at least the visual portion of the applications can be presented as one or more screens of display, the method comprising the steps of:

a. structuring applications so that they may be presented, through the network, at a first portion of one or more screens of display; and

b. structuring advertising in a manner compatible to that of the applications so that it may be presented, through the network, at a second portion of one or more screens of display concurrently with applications, wherein structuring the advertising includes configuring the advertising as objects that include advertising data and;

c. selectively storing advertising objects at a store established at the reception system.

2. The method of claim 1 wherein storing advertising objects at the reception system includes replenishing the store of advertising objects from the network when the store of advertising objects falls below a predetermined level.

3. The method of claim 2 wherein storing advertising objects at the reception system includes storing advertising

40

object identifications at the reception system, and wherein the storing of advertising object identification is based on an establishing of a characterization for the respective reception system users.

5. The method of claim 3 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on the applications requested by the respective users.

6. The method of claim 3 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on the demographic data for the respective users.

7. The method of claim 3 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on data concerning the geographical location of the respective user's reception system.

8. A method for presenting advertising in a computer network, the network including a multiplicity of user reception systems at which respective users can request applications that include interactive services, the method comprising the steps of:

a. compiling data concerning the respective users;

b. establishing characterizations for respective users based on the compiled data; and

c. structuring advertising so that it may be selectively supplied to and retrieved at the reception systems for presentation to the respective users in accordance with the characterizations established for the respective reception system users, wherein structuring advertising includes supplying advertising data to the reception system and storing a predetermined amount of the advertising data in a store established at the respective reception systems.

9. The method of claim 8 wherein supplying advertising data to the reception system includes pre-fetching advertising data from the network when the store of advertising data falls below a predetermined level.

10. The method of claim 9 wherein pre-fetching advertising data is dependent on the size of the advertising data store.

11. The method of claim 10 wherein storing advertising data at the reception system includes maintaining a list identifying the advertising data to be presented.

12. The method of claim 8 wherein the supplying of advertising data to the reception system for presentation includes the reception system requesting advertising data from the network when advertising data sought to be presented is unavailable at the reception system.

13. A method for presenting advertising in a computer network, the network including a multiplicity of user reception systems at which respective users can request applications that include interactive services, the respective reception systems including a monitor at which at least the visual portion of the applications can be presented as one or more screens of display, the method comprising the steps of:

a. structuring applications so that they may be presented at a first portion of one or more screens of display;

b. configuring the advertising as objects that include advertising data,

US 7,072,849 B1

41

- c. structuring the advertising objects in a manner compatible to that of the applications so that advertising data from an advertising object may be presented at a second portion of one or more screens of display concurrently with applications, and;
- d. selectively storing advertising objects at a store established at the reception system.

14. A method for presenting advertising obtained from a computer network, the network including a multiplicity of user reception systems at which respective users can request applications from the network that include interactive services, the respective reception systems including a monitor at which at least the visual portion of the applications can be presented as one or more screens of display, the method comprising the steps of:

- a. structuring applications so that a user requested application may be presented, through the network, at a first portion of one or more screens of display;
- b. separately structuring the advertising in a manner compatible to that of the applications so that advertising may be presented, through the network, at a second portion of one or more screens of display concurrently with any one of a plurality of user requested applications,
- c. configuring the advertising as objects that include advertising data, and
- d. selectively storing advertising objects at a store established at the reception system.

15. The method of claim 14 wherein storing advertising objects at the reception system includes replenishing the store of advertising objects from the network when the store of advertising objects falls below a predetermined level.

16. The method of claim 15 wherein storing advertising objects at the reception system includes storing advertising object identifications at the reception system, and wherein the storing of advertising object identification is based on an establishing of a characterization for the respective reception system users.

17. The method of claim 16 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on the applications requested by the respective users.

18. The method of claim 16 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on the demographic data for the respective users.

19. The method of claim 16 wherein establishing the characterization for the respective reception system users

42

includes basing the characterization at least in part on data concerning the geographical location of the respective user's reception system.

20. The method of claim 16 wherein establishing the characterization for the respective reception system users includes basing the characterization at least in part on a combination of data concerning user application requests, user demographics and geographical location of the respective user's reception system.

21. A method for presenting advertising obtained from a computer network, the network including a multiplicity of user reception systems at which respective users can request, from the network, applications that include interactive services, the method comprising the steps of:

- a. compiling data concerning the respective users;
- b. establishing characterizations for respective users based on the compiled data; and
- c. structuring advertising separately from the applications so that the advertising may be selectively supplied, through the network, to and retrieved at the reception systems for presentation to the respective users along with a requested application in accordance with the characterizations established for the respective reception system users,

wherein supplying advertising data to the reception system includes storing a predetermined amount of the advertising data in a store established at the respective reception systems.

22. The method of claim 21 wherein supplying advertising data to the reception system includes pre-fetching advertising data from the network when the store of advertising data falls below a predetermined level.

23. The method of claim 22 wherein pre-fetching advertising data is dependent on the size of the advertising data store.

24. The method of claim 23 wherein storing advertising data at the reception system includes maintaining a list identifying the advertising data to be presented.

25. The method of claim 21 wherein the supplying of advertising data to the reception system for presentation includes the reception system requesting advertising data from the network when advertising data sought to be presented is unavailable at the reception system.

* * * * *

EXHIBIT B



US007631346B2

(12) **United States Patent**
Hinton et al.

(10) **Patent No.:** US 7,631,346 B2
(45) **Date of Patent:** Dec. 8, 2009

(54) **METHOD AND SYSTEM FOR A RUNTIME USER ACCOUNT CREATION OPERATION WITHIN A SINGLE-SIGN-ON PROCESS IN A FEDERATED COMPUTING ENVIRONMENT**

(75) Inventors: **Heather Maria Hinton**, Austin, TX (US); **Ivan Matthew Milman**, Austin, TX (US); **Venkat Raghavan**, Austin, TX (US); **Shane Bradley Weeden**, Gold Coast (AU)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 827 days.

(21) Appl. No.: **11/097,587**

(22) Filed: **Apr. 1, 2005**

(65) **Prior Publication Data**

US 2006/0236382 A1 Oct. 19, 2006

(51) **Int. Cl.**

G06F 7/04 (2006.01)
G06F 15/16 (2006.01)
G04L 9/32 (2006.01)
G06F 17/30 (2006.01)

(52) **U.S. Cl.** **726/8; 380/279**

(58) **Field of Classification Search** **726/6, 726/5, 4, 8; 709/223, 229, 219; 713/202, 713/186, 169; 380/279; 715/500**

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

- 7,290,278 B2 * 10/2007 Cahill et al. 726/6
2003/0149781 A1 8/2003 Yared et al. 709/223
2003/0154266 A1 * 8/2003 Bobick et al. 709/229
2004/0010607 A1 * 1/2004 Lee et al. 709/229
2004/0158746 A1 * 8/2004 Hu et al. 713/202

2004/0205176 A1 * 10/2004 Ting et al. 709/223
2005/0074126 A1 * 4/2005 Stanko 380/279
2005/0210270 A1 * 9/2005 Rohatgi et al. 713/186
2005/0240763 A9 * 10/2005 Bhat et al. 713/169
2005/0257130 A1 * 11/2005 Ito 715/500.1
2006/0048213 A1 * 3/2006 Cheng et al. 726/5

(Continued)

OTHER PUBLICATIONS

Gross, T.; Security analysis of the SAML single sign-on browser/artifact profile; Publication Date: Dec. 8-12, 2003; IBM Zurich Res. Lab; On pp. 298-307.*

Primary Examiner—Kambiz Zand

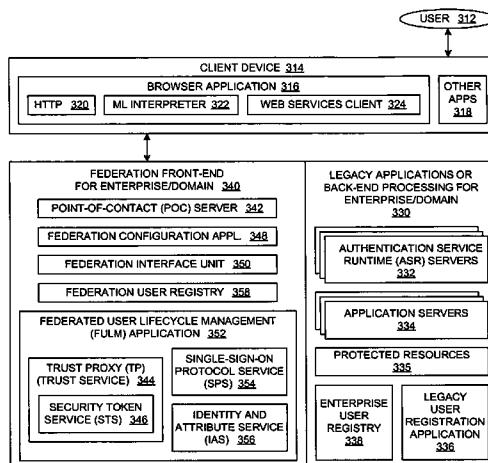
Assistant Examiner—Monjour Rahim

(74) *Attorney, Agent, or Firm*—Jeffrey S. LaBaw; David H. Judson

(57) **ABSTRACT**

A method, system, apparatus, and computer program product are presented to support computing systems of different enterprises that interact within a federated computing environment. Federated single-sign-on operations can be initiated at the computing systems of federation partners on behalf of a user even though the user has not established a user account at a federation partner prior to the initiation of the single-sign-on operation. For example, an identity provider can initiate a single-sign-on operation at a service provider while attempting to obtain access to a controlled resource on behalf of a user. When the service provider recognizes that it does not have a linked user account for the user that allows for a single-sign-on operation with the identity provider, the service provider creates a local user account. The service provider can also pull user attributes from the identity provider as necessary to perform the user account creation operation.

20 Claims, 14 Drawing Sheets



US 7,631,346 B2

Page 2

U.S. PATENT DOCUMENTS

2006/0059544 A1 * 3/2006 Guthrie et al. 726/4
2006/0195893 A1 * 8/2006 Caceres et al. 726/8

2007/0005730 A1 * 1/2007 Torvinen et al. 709/219

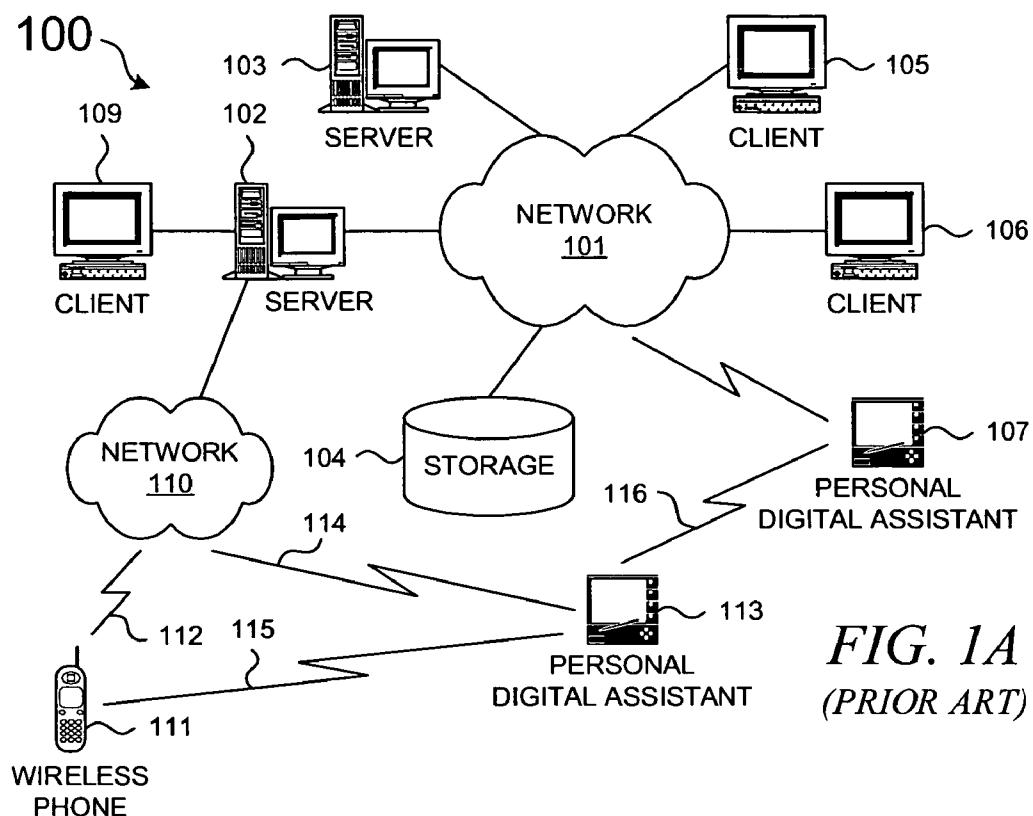
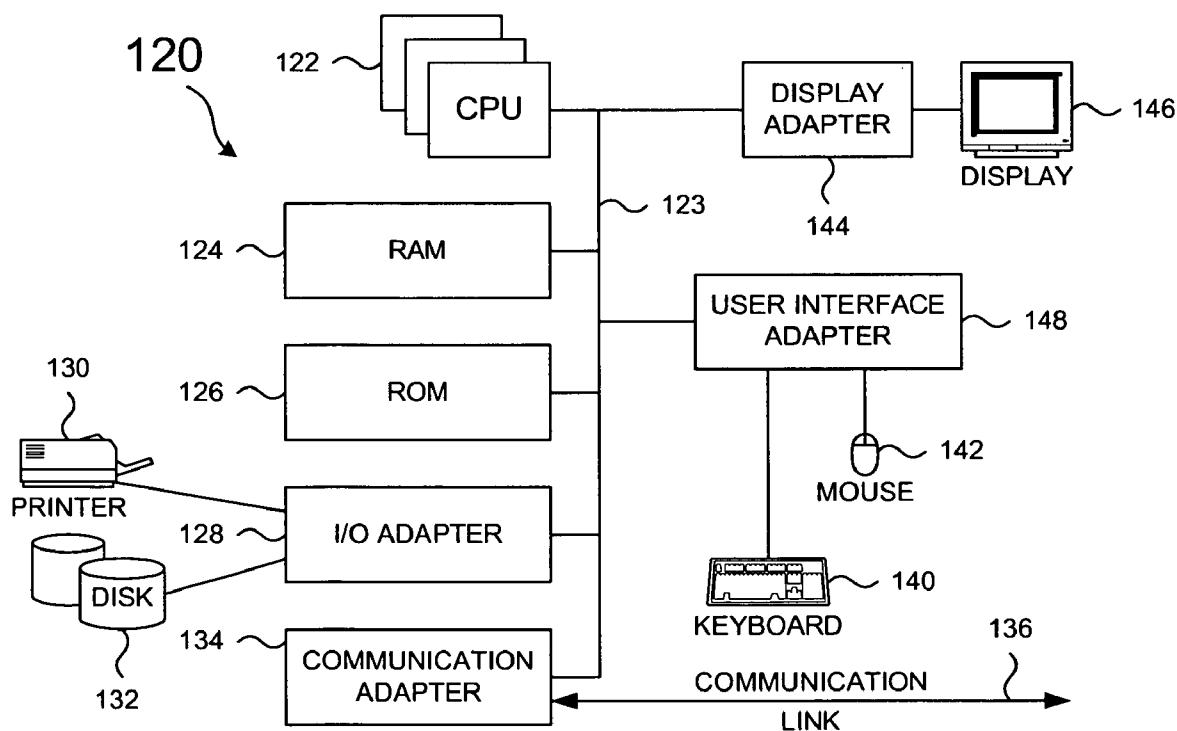
* cited by examiner

U.S. Patent

Dec. 8, 2009

Sheet 1 of 14

US 7,631,346 B2

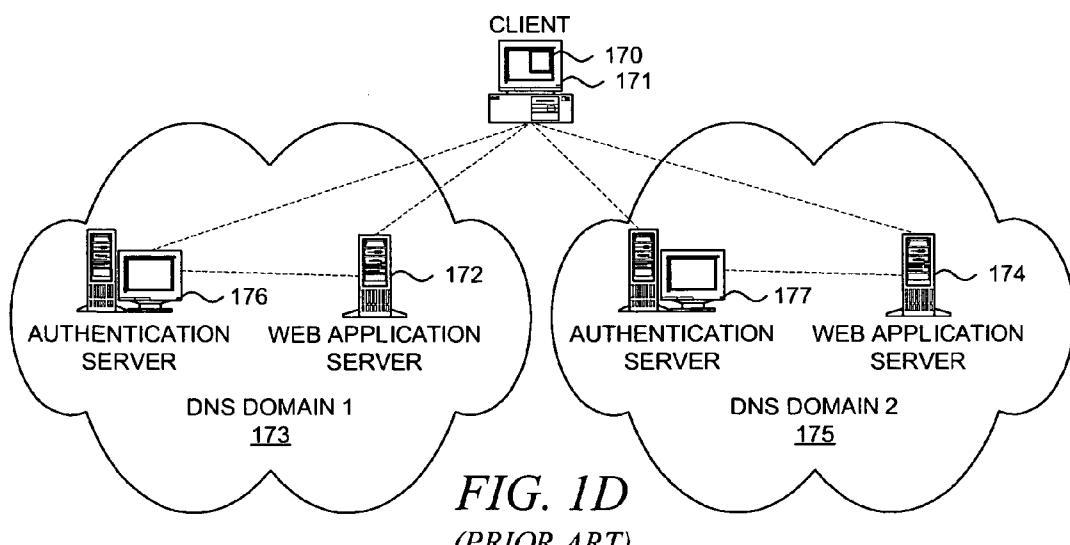
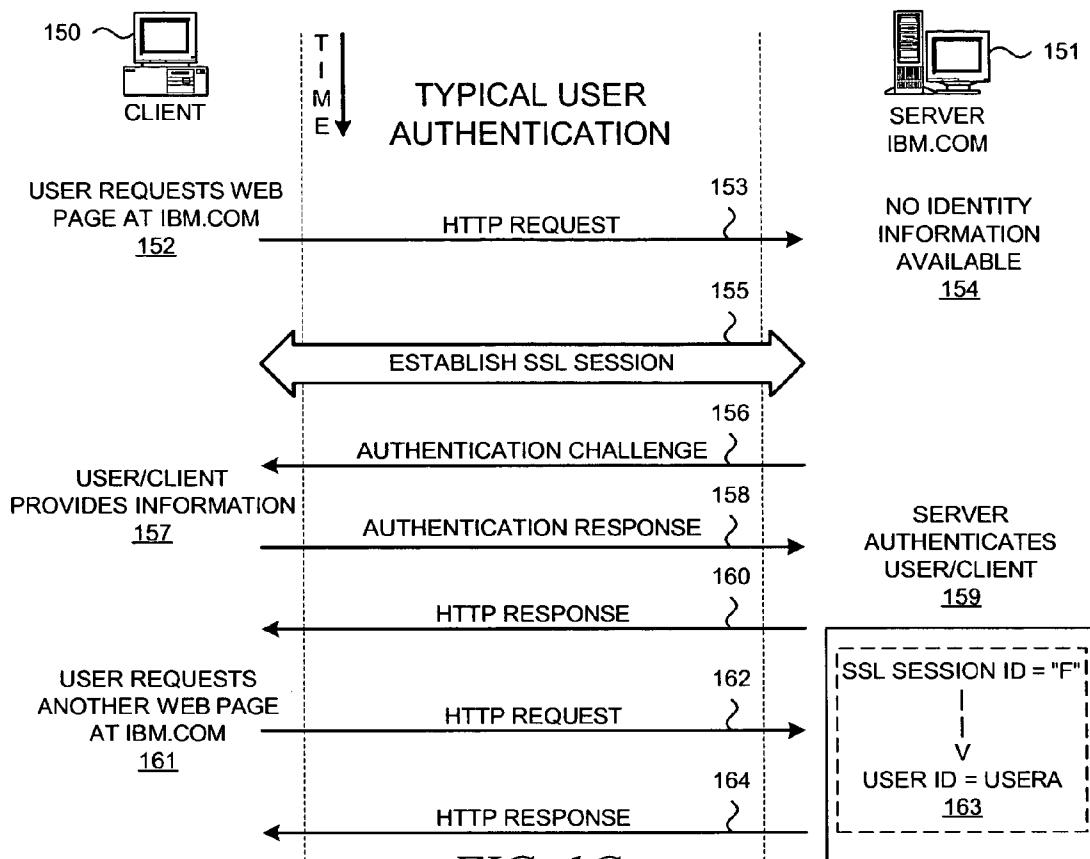
FIG. 1A
(PRIOR ART)FIG. 1B
(PRIOR ART)

U.S. Patent

Dec. 8, 2009

Sheet 2 of 14

US 7,631,346 B2

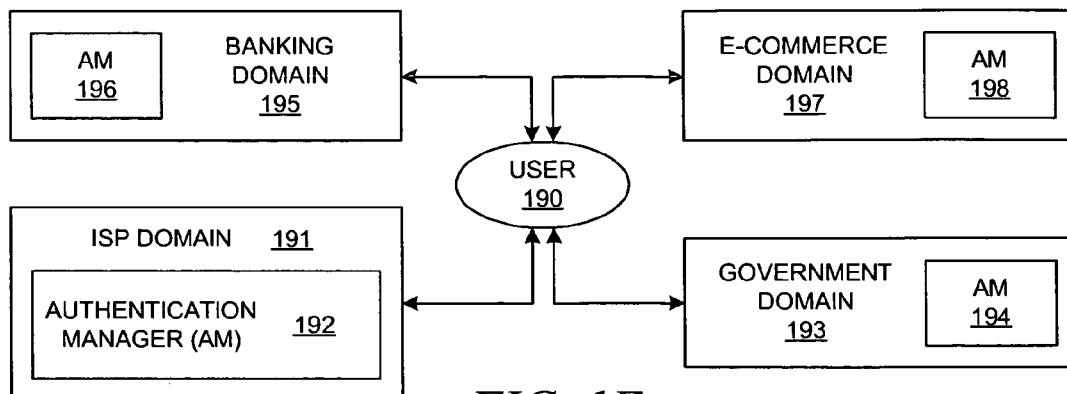
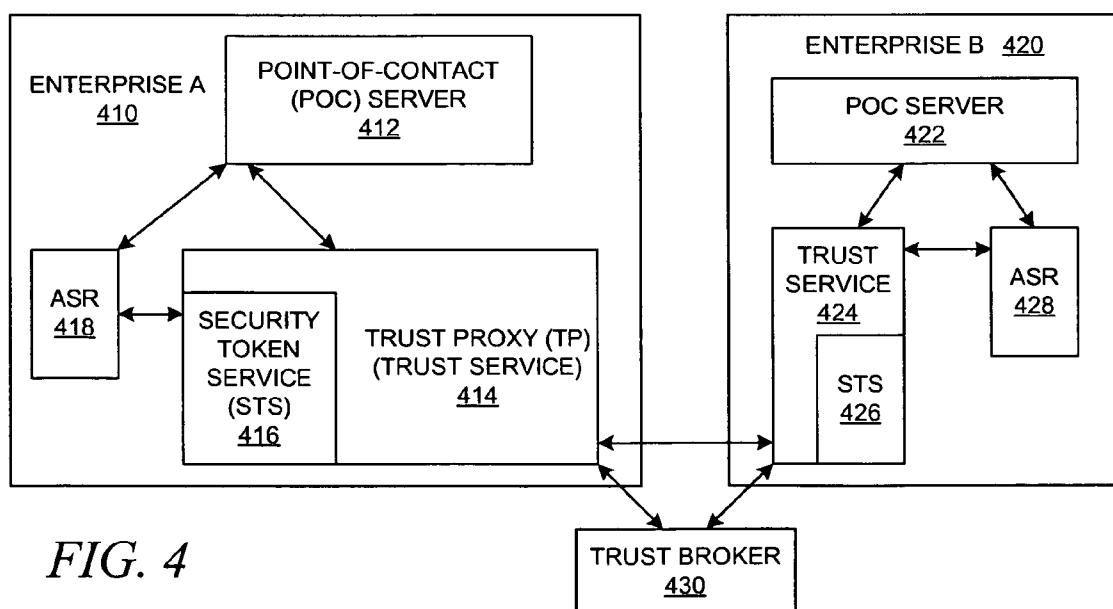
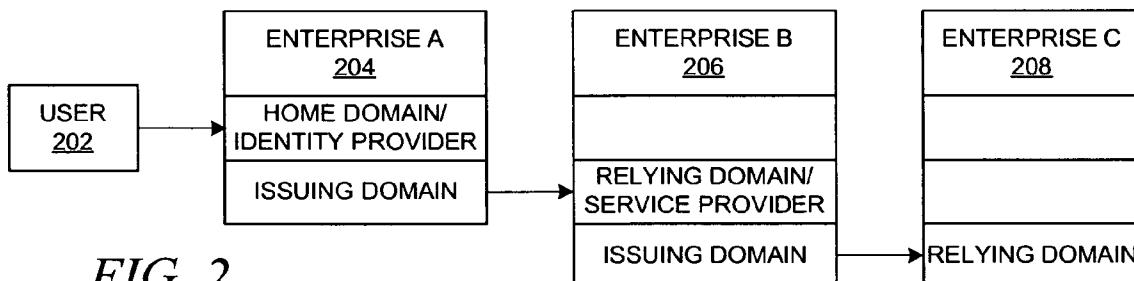


U.S. Patent

Dec. 8, 2009

Sheet 3 of 14

US 7,631,346 B2

FIG. 1E
(PRIOR ART)

U.S. Patent

Dec. 8, 2009

Sheet 4 of 14

US 7,631,346 B2

FIG. 3

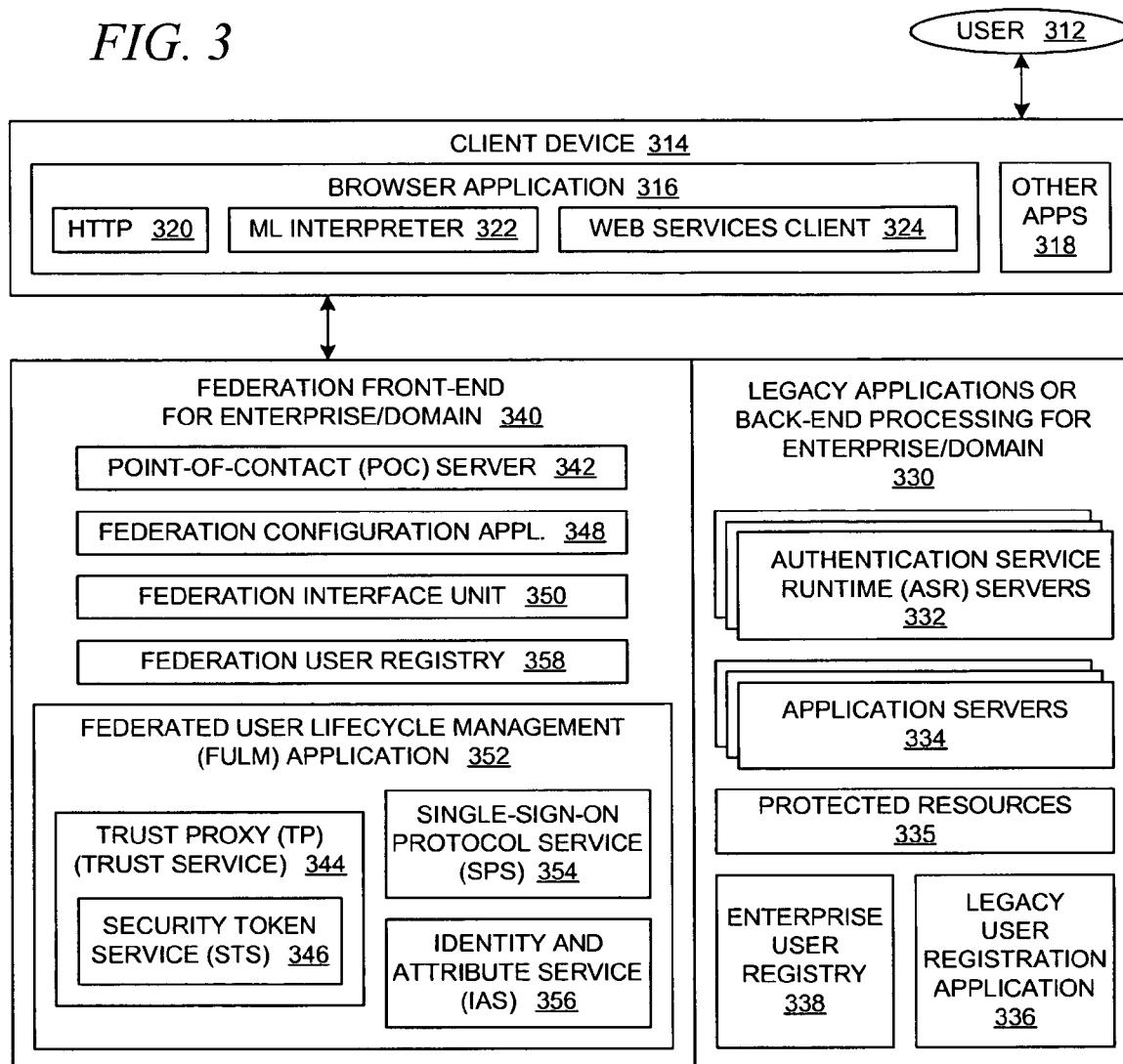
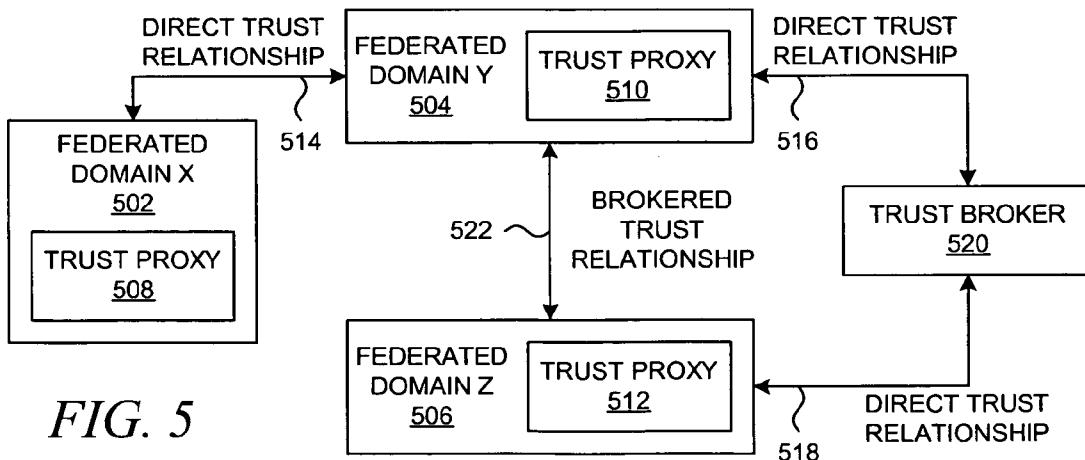


FIG. 5

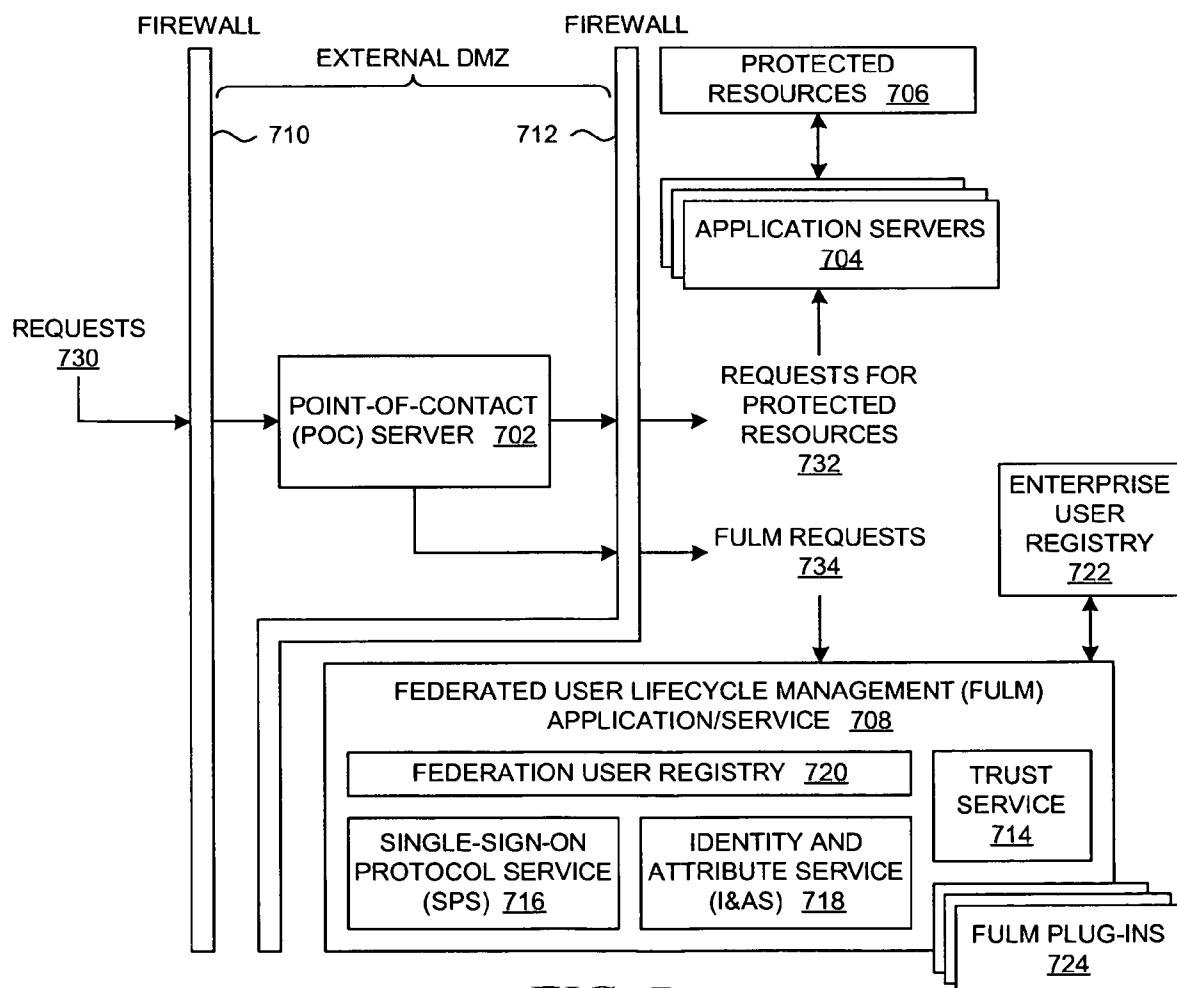
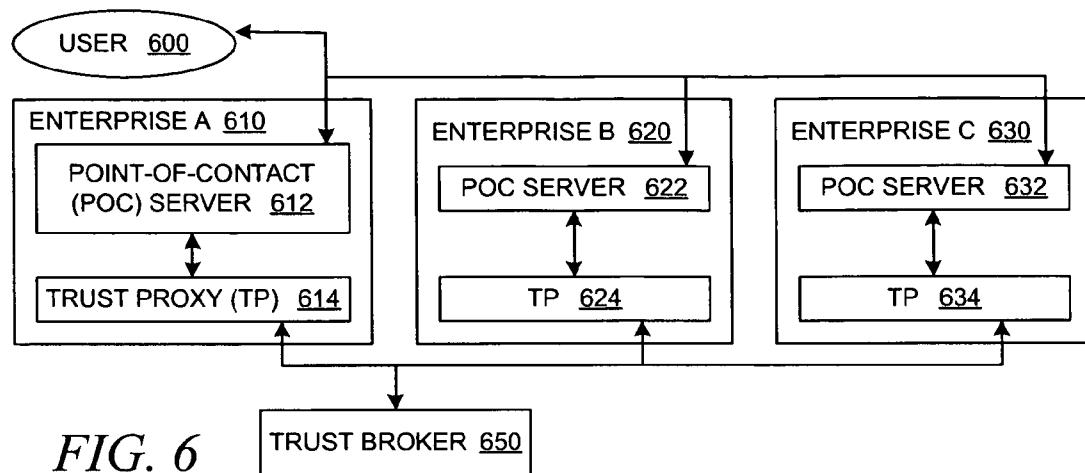


U.S. Patent

Dec. 8, 2009

Sheet 5 of 14

US 7,631,346 B2

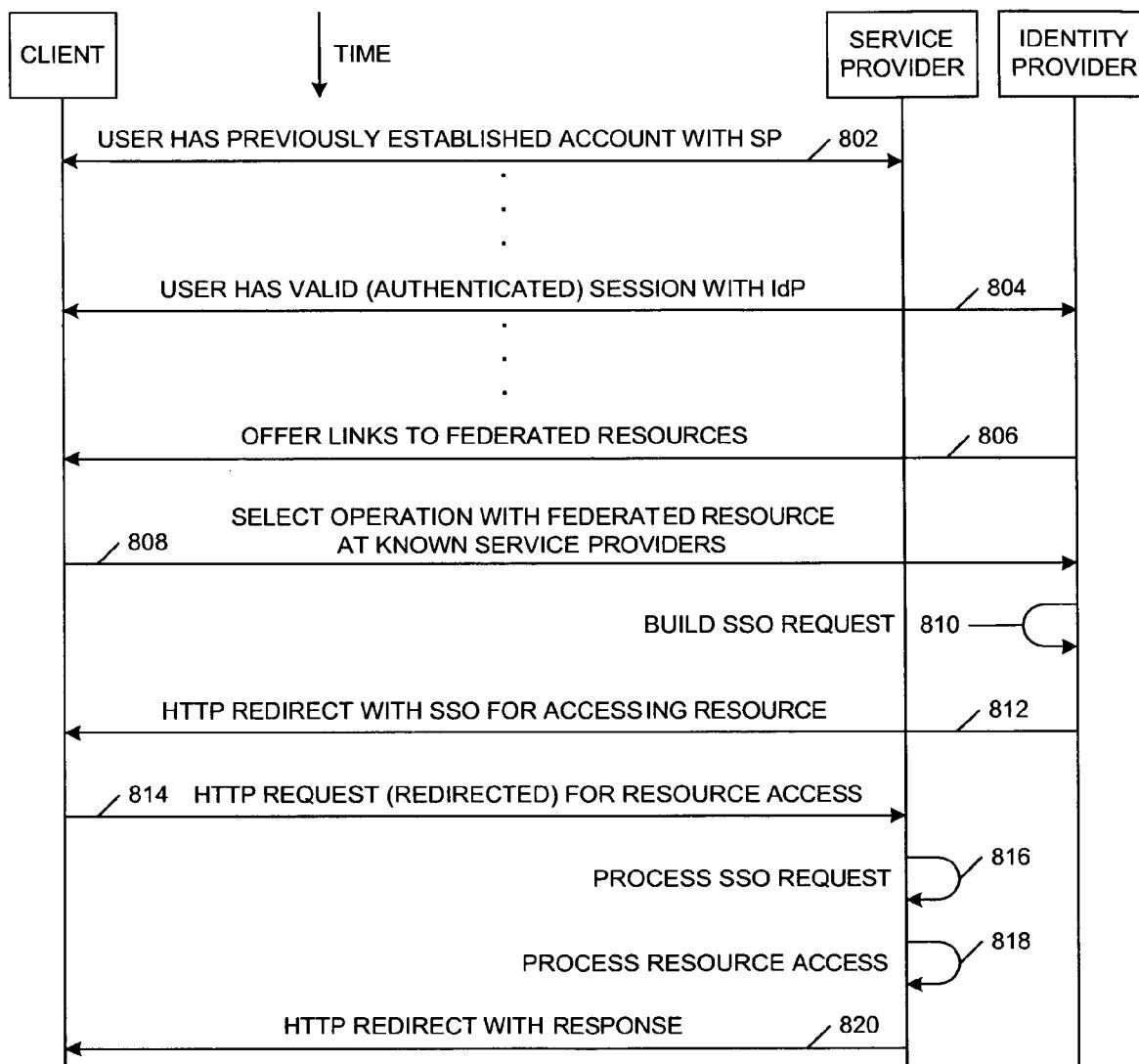


U.S. Patent

Dec. 8, 2009

Sheet 6 of 14

US 7,631,346 B2



**TYPICAL SINGLE-SIGN-ON OPERATION
(INITIATED BY IDENTITY PROVIDER -- USER PREVIOUSLY PROVISIONED AT SP)**

FIG. 8
(PRIOR ART)

U.S. Patent

Dec. 8, 2009

Sheet 7 of 14

US 7,631,346 B2

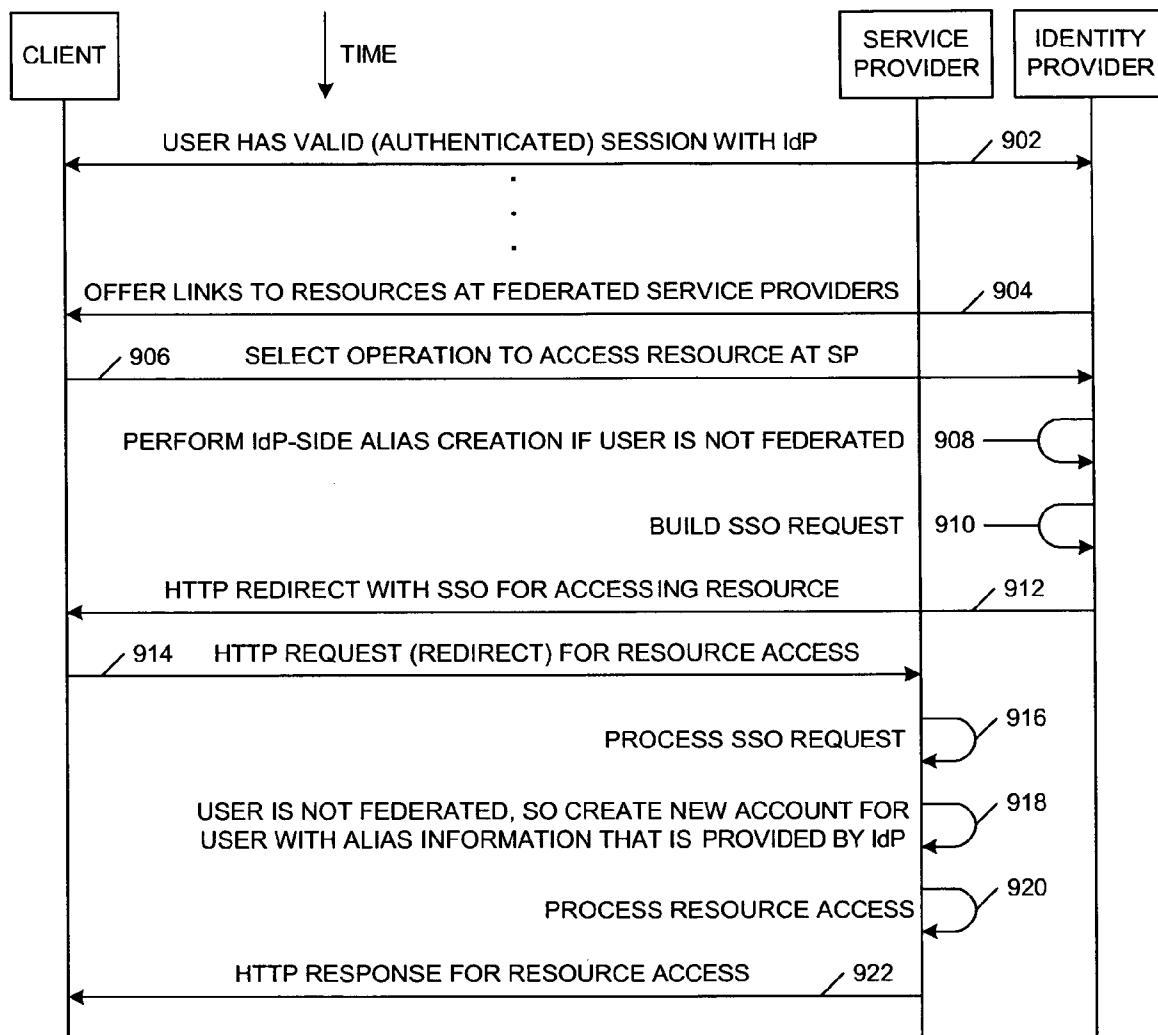


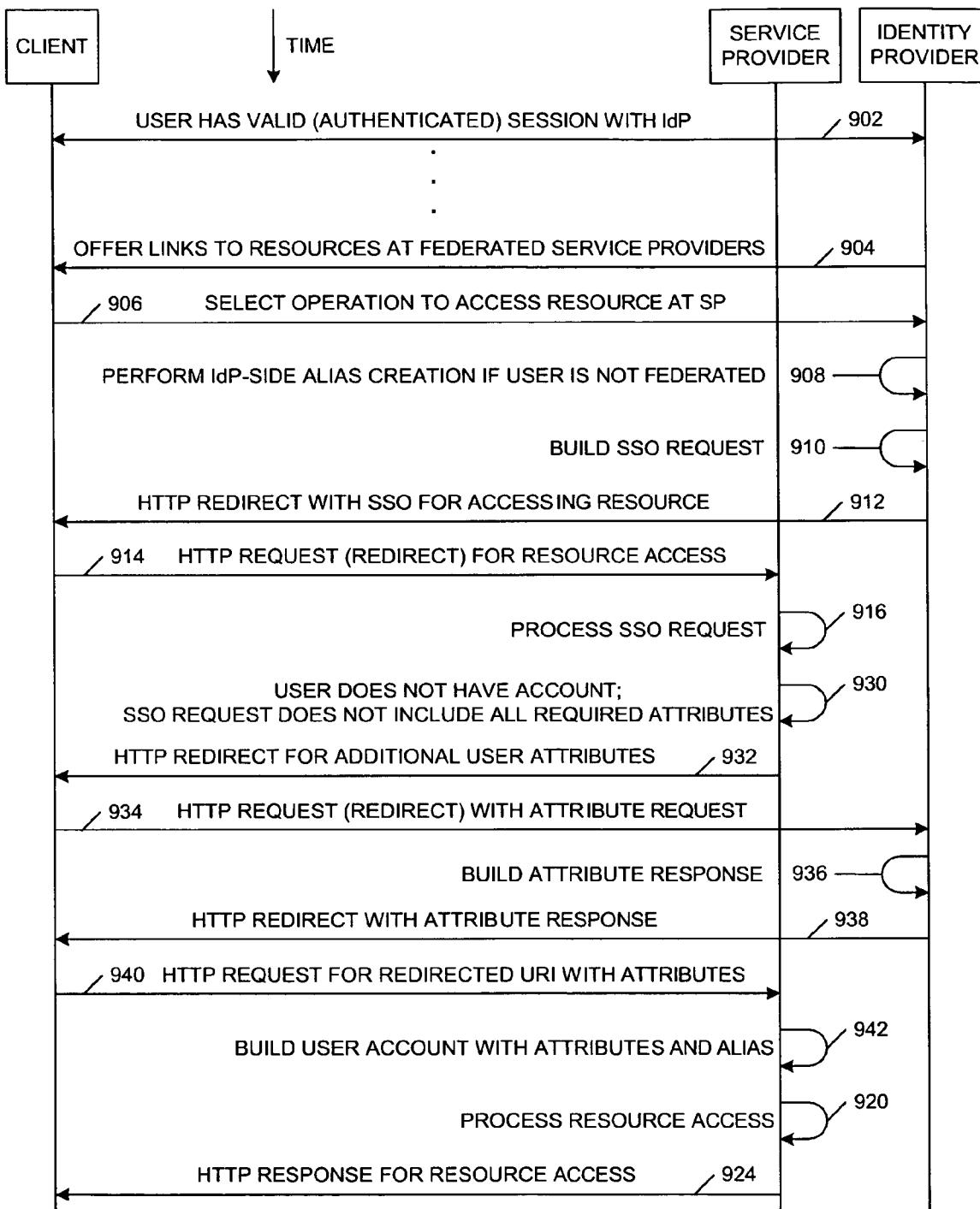
FIG. 9A

U.S. Patent

Dec. 8, 2009

Sheet 8 of 14

US 7,631,346 B2



PUSH-TYPE SINGLE-SIGN-ON OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP (ADDITIONAL PULLING OF USER ATTRIBUTES BY SP FROM IDP)

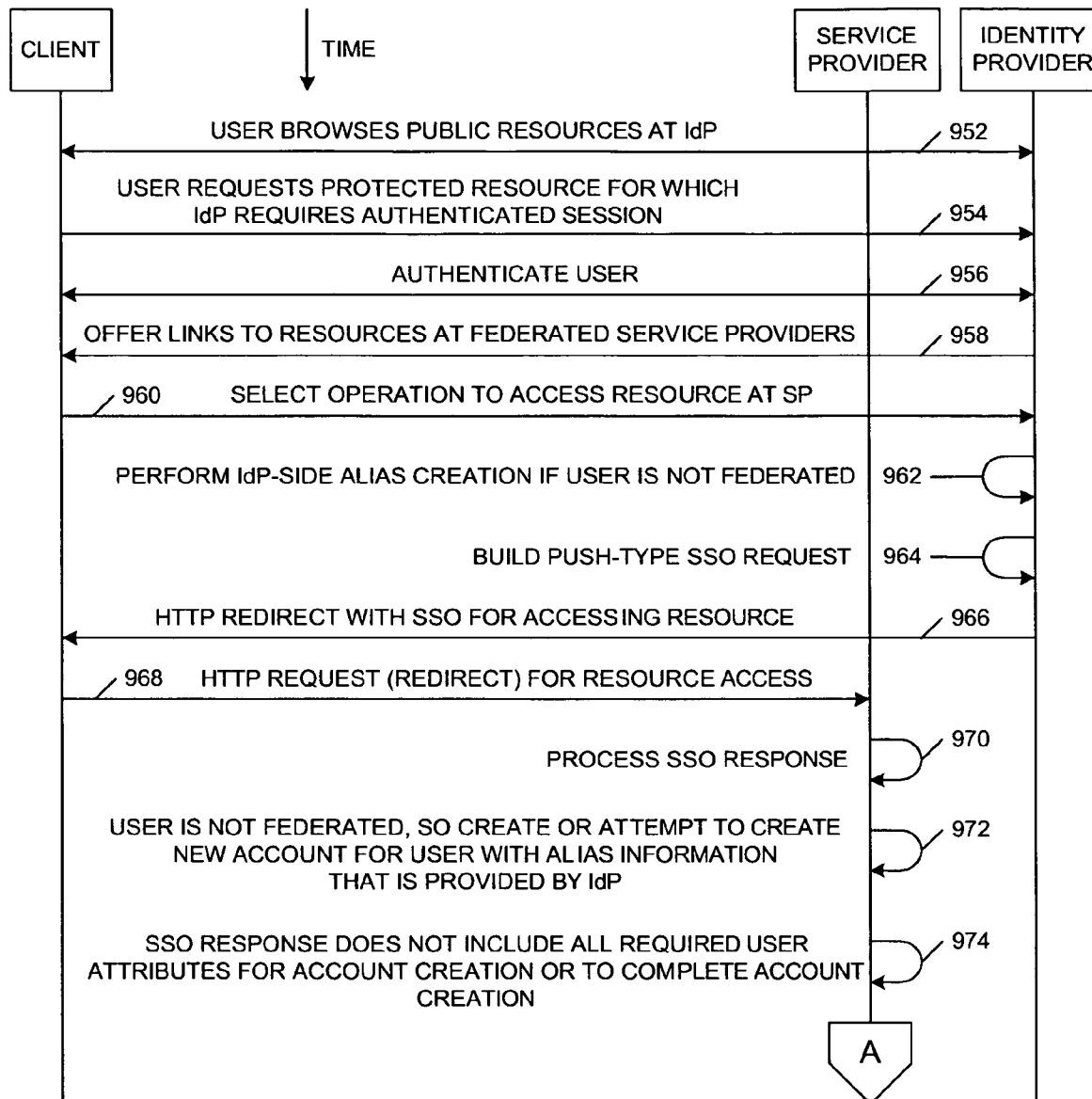
FIG. 9B

U.S. Patent

Dec. 8, 2009

Sheet 9 of 14

US 7,631,346 B2



PUSH-TYPE SINGLE-SIGN-ON OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP
(ADDITIONAL PULLING OF USER ATTRIBUTES BY SP FROM IDP)

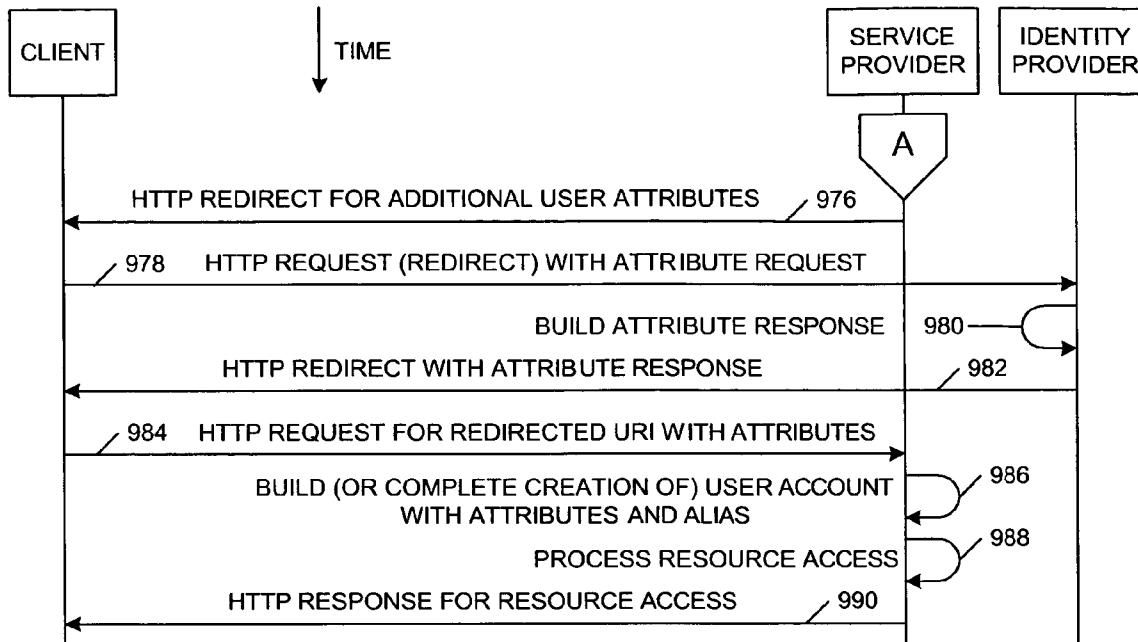
FIG. 9C

U.S. Patent

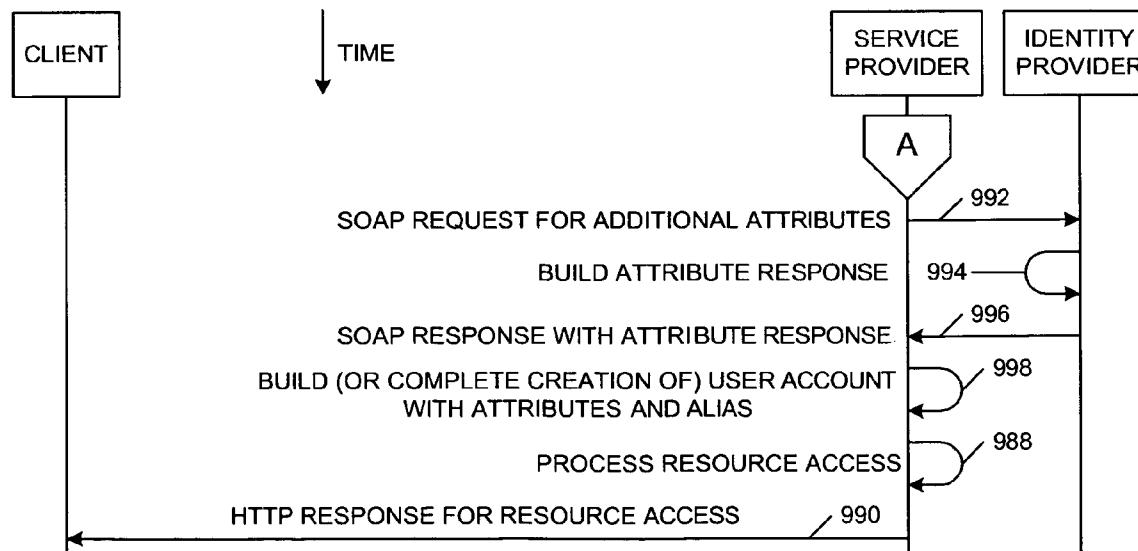
Dec. 8, 2009

Sheet 10 of 14

US 7,631,346 B2



**COMPLETION OF PUSH-TYPE SSO OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP
(FRONT-CHANNEL USER ATTRIBUTE RETRIEVAL BY SP FROM IDP)**

FIG. 9D

**COMPLETION OF PUSH-TYPE SSO OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP
(BACK-CHANNEL USER ATTRIBUTE RETRIEVAL BY SP FROM IDP)**

FIG. 9E

U.S. Patent

Dec. 8, 2009

Sheet 11 of 14

US 7,631,346 B2

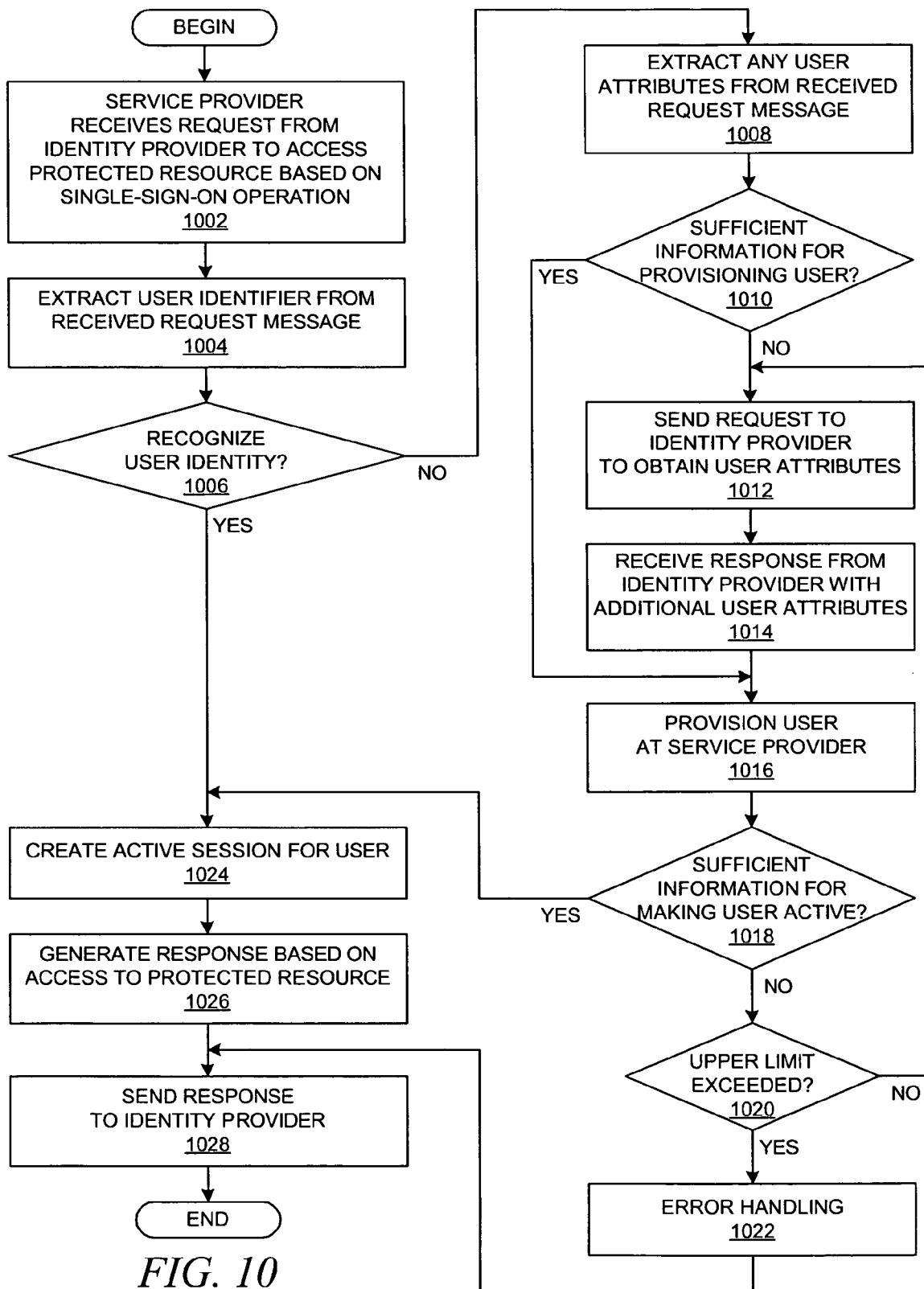


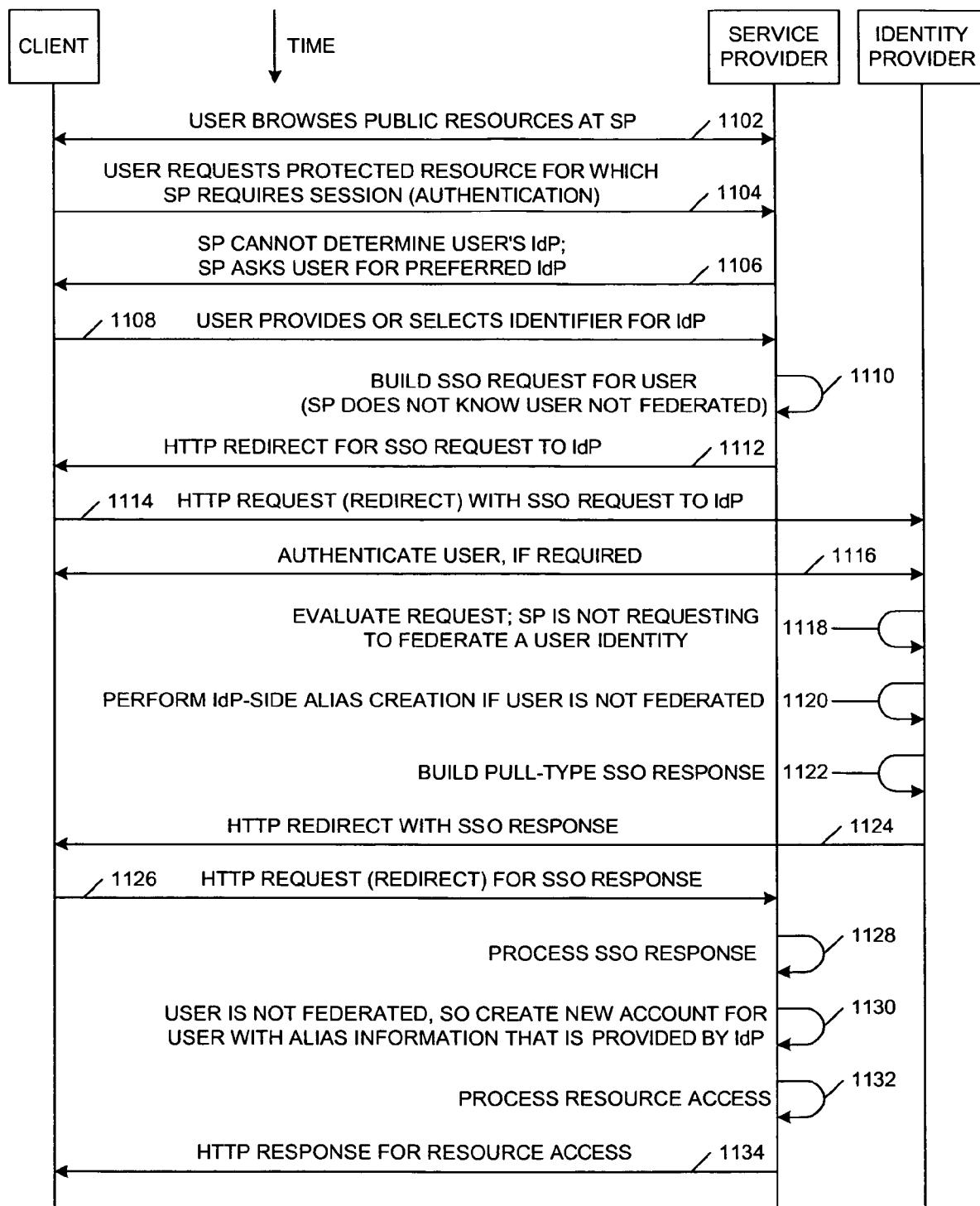
FIG. 10

U.S. Patent

Dec. 8, 2009

Sheet 12 of 14

US 7,631,346 B2



PULL-TYPE SINGLE-SIGN-ON OPERATION WITH RUNTIME USER PROVISIONING AT SP
(USER NOT PREVIOUSLY PROVISIONED AT SP)

FIG. 11A

U.S. Patent

Dec. 8, 2009

Sheet 13 of 14

US 7,631,346 B2

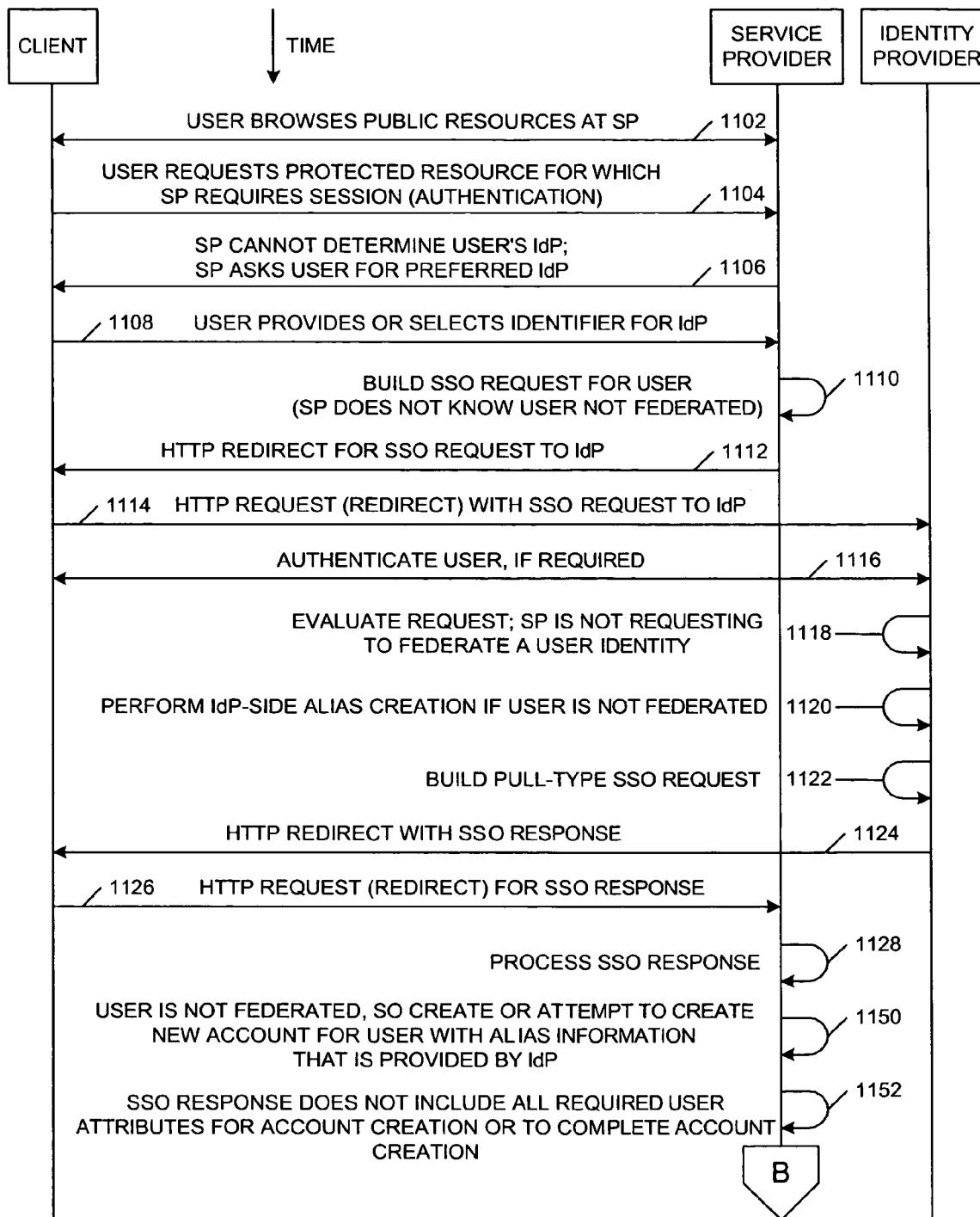


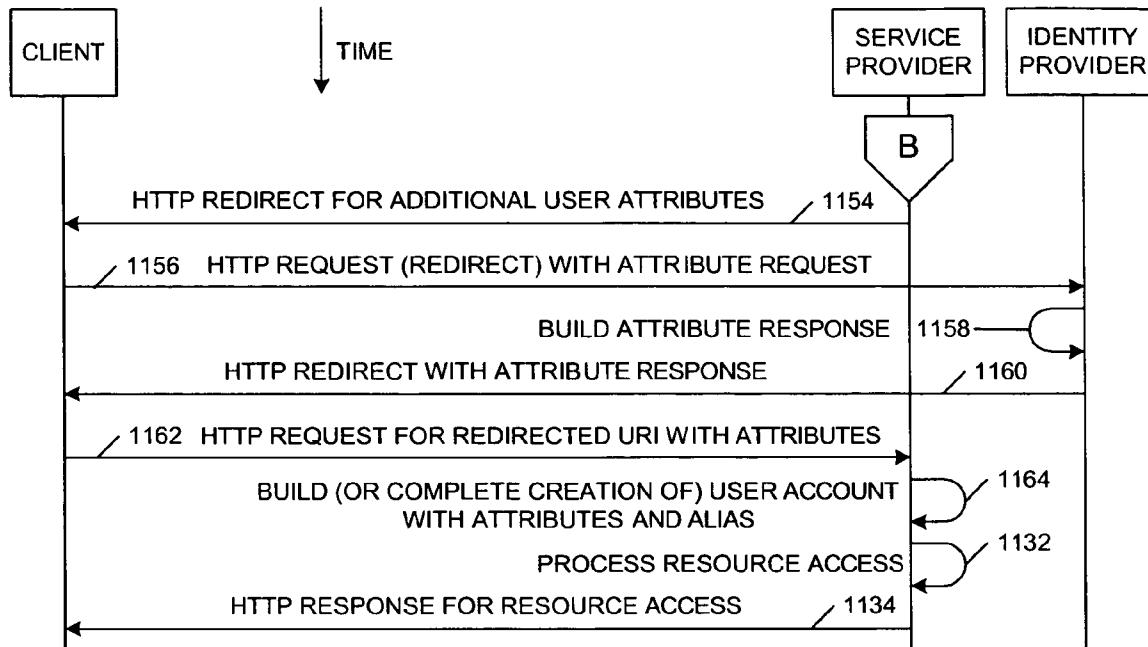
FIG. 11B

U.S. Patent

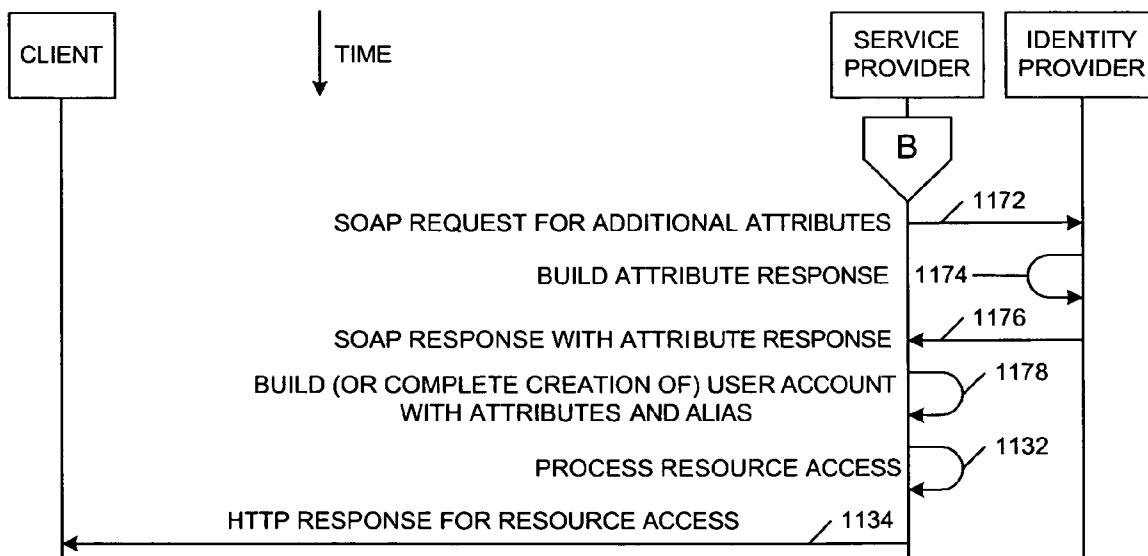
Dec. 8, 2009

Sheet 14 of 14

US 7,631,346 B2



**COMPLETION OF PULL-TYPE SSO OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP
(FRONT-CHANNEL USER ATTRIBUTE RETRIEVAL BY SP FROM IDP)**

FIG. 11C

**COMPLETION OF PULL-TYPE SSO OPERATION WITH RUNTIME USER ACCOUNT CREATION AT SP
(BACK-CHANNEL USER ATTRIBUTE RETRIEVAL BY SP FROM IDP)**

FIG. 11D

US 7,631,346 B2

1

**METHOD AND SYSTEM FOR A RUNTIME
USER ACCOUNT CREATION OPERATION
WITHIN A SINGLE-SIGN-ON PROCESS IN A
FEDERATED COMPUTING ENVIRONMENT**

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to an improved data processing system and, in particular, to a method and apparatus for multicommputer data transferring. Still more particularly, the present invention is directed to networked computer systems.

2. Description of Related Art

Enterprises generally desire to provide authorized users with secure access to protected resources in a user-friendly manner throughout a variety of networks, including the Internet. Although providing secure authentication mechanisms reduces the risks of unauthorized access to protected resources, those authentication mechanisms may become barriers to accessing protected resources. Users generally desire the ability to change from interacting with one application to another application without regard to authentication barriers that protect each particular system supporting those applications.

As users get more sophisticated, they expect that computer systems coordinate their actions so that burdens on the user are reduced. These types of expectations also apply to authentication processes. A user might assume that once he or she has been authenticated by some computer system, the authentication should be valid throughout the user's working session, or at least for a particular period of time, without regard to the various computer architecture boundaries that are almost invisible to the user. Enterprises generally try to fulfill these expectations in the operational characteristics of their deployed systems, not only to placate users but also to increase user efficiency, whether the user efficiency is related to employee productivity or customer satisfaction.

More specifically, with the current computing environment in which many applications have a Web-based user interface that is accessible through a common browser, users expect more user-friendliness and low or infrequent barriers to movement from one Web-based application to another. In this context, users are coming to expect the ability to jump from interacting with an application on one Internet domain to another application on another domain without regard to the authentication barriers that protect each particular domain. However, even if many systems provide secure authentication through easy-to-use, Web-based interfaces, a user may still be forced to reckon with multiple authentication processes that stymie user access across a set of domains. Subjecting a user to multiple authentication processes in a given time frame may significantly affect the user's efficiency.

For example, various techniques have been used to reduce authentication burdens on users and computer system administrators. These techniques are generally described as "single-sign-on" (SSO) processes because they have a common purpose: after a user has completed a sign-on operation, i.e. been authenticated, the user is subsequently not required to perform another authentication operation. Hence, the goal is that the user would be required to complete only one authentication process during a particular user session.

To reduce the costs of user management and to improve interoperability among enterprises, federated computing spaces have been created. A federation is a loosely coupled affiliation of enterprises which adhere to certain standards of interoperability; the federation provides a mechanism for trust among those enterprises with respect to certain compu-

2

tational operations for the users within the federation. For example, a federation partner may act as a user's home domain or identity provider. Other partners within the same federation may rely on the user's identity provider for primary management of the user's authentication credentials, e.g., accepting a single-sign-on token that is provided by the user's identity provider.

As enterprises move to support federated business interactions, these enterprises should provide a user experience that reflects the increased cooperation between two businesses. As noted above, a user may authenticate to one party that acts as an identity provider and then single-sign-on to a federated business partner that acts as a service provider. In conjunction with single-sign-on functionality, additional user lifecycle functionality, such as single-sign-off, user provisioning, and account linking/delinking, should also be supported.

Single-sign-on solutions require that a user be identifiable in some form or another at both an identity provider and a service provider; the identity provider needs to be able to identify and authenticate a user, and the service provider needs to be able to identify the user based on some form of assertion about the user in response to a single-sign-on request. Various prior art single-sign-on solutions, e.g., such as those described in the Liberty Alliance ID-FF specifications, require that a user have an authenticatable account at both an identity provider and a service provider as a prerequisite to a federated single-sign-on operation. Some federated solutions support an a priori user account creation event across domains to be used to establish these accounts, thereby satisfying a requirement that a user have an authenticatable account at both an identity provider and a service provider as a prerequisite to a federated single-sign-on operation. Although some federated solutions provide a robust set of federated user lifecycle management operations, such as user account creation, user account management, user attribute management, account suspension, and account deletion, these federated management systems do not provide a lightweight solution that is suitable for certain federation partners or for certain federated purposes.

Therefore, it would be advantageous to have methods and systems in which enterprises can provide comprehensive single-sign-on experiences to users in a federated computing environment in a lightweight manner that does not require an extensive amount of a priori processing.

SUMMARY OF THE INVENTION

A method, system, apparatus, and computer program product are presented to support computing systems of different enterprises that interact within a federated computing environment. Federated single-sign-on operations can be initiated at the computing systems of federation partners on behalf of a user even though the user has not established a user account at a federation partner prior to the initiation of the single-sign-on operation. For example, an identity provider can initiate a single-sign-on operation at a service provider while attempting to obtain access to a controlled resource on behalf of a user. When the service provider recognizes that it does not have a linked user account for the user that allows a single-sign-on operation from the identity provider, the service provider creates a local user account based at least in part on information from the identity provider. The service provider

US 7,631,346 B2

3

can also pull user attributes from the identity provider as necessary to perform the user account creation operation.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features believed characteristic of the invention are set forth in the appended claims. The invention itself, further objectives, and advantages thereof, will be best understood by reference to the following detailed description when read in conjunction with the accompanying drawings, wherein:

FIG. 1A depicts a typical network of data processing systems, each of which may implement the present invention;

FIG. 1B depicts a typical computer architecture that may be used within a data processing system in which the present invention may be implemented;

FIG. 1C depicts a data flow diagram that illustrates a typical authentication process that may be used when a client attempts to access a protected resource at a server;

FIG. 1D depicts a network diagram that illustrates a typical Web-based environment in which the present invention may be implemented;

FIG. 1E depicts a block diagram that illustrates an example of a typical online transaction that might require multiple authentication operations from a user;

FIG. 2 depicts a block diagram that illustrates the terminology of the federated environment with respect to a transaction that is initiated by a user to a first federated enterprise, which, in response, invokes actions at downstream entities within the federated environment;

FIG. 3 depicts a block diagram that illustrates the integration of pre-existing data processing systems at a given domain with some federated architecture components that may be used to support an embodiment of the present invention;

FIG. 4 depicts a block diagram that illustrates an example of a manner in which some components within a federated architecture may be used to establish trust relationships to support an implementation of the present invention;

FIG. 5 depicts a block diagram that illustrates an exemplary set of trust relationships between federated domains using trust proxies and a trust broker in accordance with an exemplary federated architecture that is able to support the present invention;

FIG. 6 depicts a block diagram that illustrates a federated environment that supports federated single-sign-on operations;

FIG. 7 depicts a block diagram that illustrates some of the components in a federated domain for implementing federated user lifecycle management functionality in order to support the present invention;

FIG. 8 depicts a dataflow diagram that shows a typical prior art HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider;

FIGS. 9A-9B depicts dataflow diagrams that show an HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention;

FIGS. 9C-9E depict dataflow diagrams that show an HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider with alter-

4

native methods for obtaining user attributes by the federated service provider in accordance with an embodiment of the present invention;

FIG. 10 depicts a flowchart that shows a more detailed process for performing a runtime linked-user-account creation operation at a service provider during a single-sign-on operation that has been initiated by an identity provider;

FIG. 11A depicts a dataflow diagram that shows an HTTP-redirection-based pull-type single-sign-on operation that is initiated by a federated service provider to allow access to a protected resource at the federated service provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention; and

FIGS. 11B-11D depict a set of dataflow diagrams that show an HTTP-redirection-based pull-type single-sign-on operation that is initiated by a federated service provider to allow access to a protected resource at the federated service provider with additional retrieval of user attribute information from a federated identity provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

In general, the devices that may comprise or relate to the present invention include a wide variety of data processing technology. Therefore, as background, a typical organization of hardware and software components within a distributed data processing system is described prior to describing the present invention in more detail.

With reference now to the figures, FIG. 1A depicts a typical network of data processing systems, each of which may implement the present invention. Distributed data processing system 100 contains network 101, which is a medium that may be used to provide communications links between various devices and computers connected together within distributed data processing system 100. Network 101 may include permanent connections, such as wire or fiber optic cables, or temporary connections made through telephone or wireless communications. In the depicted example, server 102 and server 103 are connected to network 101 along with storage unit 104. In addition, clients 105-107 also are connected to network 101. Clients 105-107 and servers 102-103 may be represented by a variety of computing devices, such as mainframes, personal computers, personal digital assistants (PDAs), etc. Distributed data processing system 100 may include additional servers, clients, routers, other devices, and peer-to-peer architectures that are not shown.

In the depicted example, distributed data processing system 100 may include the Internet with network 101 representing a worldwide collection of networks and gateways that use various protocols to communicate with one another, such as LDAP (Lightweight Directory Access Protocol), TCP/IP (Transport Control Protocol/Internet Protocol), HTTP (HyperText Transport Protocol), etc. Of course, distributed data processing system 100 may also include a number of different types of networks, such as, for example, an intranet, a local area network (LAN), or a wide area network (WAN). For example, server 102 directly supports client 109 and network 110, which incorporates wireless communication links. Network-enabled phone 111 connects to network 110 through wireless link 112, and PDA 113 connects to network 110 through wireless link 114. Phone 111 and PDA 113 can also directly transfer data between themselves across wireless link 115 using an appropriate technology, such as Bluetooth™.

US 7,631,346 B2

5

wireless technology, to create so-called personal area networks or personal ad-hoc networks. In a similar manner, PDA 113 can transfer data to PDA 107 via wireless communication link 116.

The present invention could be implemented on a variety of hardware platforms and software environments. FIG. 1A is intended as an example of a heterogeneous computing environment and not as an architectural limitation for the present invention.

With reference now to FIG. 1B, a diagram depicts a typical computer architecture of a data processing system, such as those shown in FIG. 1A, in which the present invention may be implemented. Data processing system 120 contains one or more central processing units (CPUs) 122 connected to internal system bus 123, which interconnects random access memory (RAM) 124, read-only memory 126, and input/output adapter 128, which supports various I/O devices, such as printer 130, disk units 132, or other devices not shown, such as a audio output system, etc. System bus 123 also connects communication adapter 134 that provides access to communication link 136. User interface adapter 148 connects various user devices, such as keyboard 140 and mouse 142, or other devices not shown, such as a touch screen, stylus, microphone, etc. Display adapter 144 connects system bus 123 to display device 146.

Those of ordinary skill in the art will appreciate that the hardware in FIG. 1B may vary depending on the system implementation. For example, the system may have one or more processors, such as an Intel® Pentium®-based processor and a digital signal processor (DSP), and one or more types of volatile and non-volatile memory. Other peripheral devices may be used in addition to or in place of the hardware depicted in FIG. 1B. The depicted examples are not meant to imply architectural limitations with respect to the present invention.

In addition to being able to be implemented on a variety of hardware platforms, the present invention may be implemented in a variety of software environments. A typical operating system may be used to control program execution within each data processing system. For example, one device may run a Unix® operating system, while another device contains a simple Java® runtime environment. A representative computer platform may include a browser, which is a well known software application for accessing hypertext documents in a variety of formats, such as graphic files, word processing files, Extensible Markup Language (XML), Hypertext Markup Language (HTML), Handheld Device Markup Language (HDML), Wireless Markup Language (WML), and various other formats and types of files. It should also be noted that the distributed data processing system shown in FIG. 1A is contemplated as being fully able to support a variety of peer-to-peer subnets and peer-to-peer services.

With reference now to FIG. 1C, a data flow diagram illustrates a typical authentication process that may be used when a client attempts to access a protected resource at a server. As illustrated, the user at a client workstation 150 seeks access over a computer network to a protected resource on a server 151 through the user's web browser executing on the client workstation. A protected or controlled resource is a resource (an application, an object, a document, a page, a file, executable code, or other computational resource, communication-type resource, etc.) for which access is controlled or restricted. A protected resource is identified by a Uniform Resource Locator (URL), or more generally, a Uniform Resource Identifier (URI), that can only be accessed by an authenticated and/or authorized user. The computer network

6

may be the Internet, an intranet, or other network, as shown in FIG. 1A or FIG. 1B, and the server may be a web application server (WAS), a server application, a servlet process, or the like.

5 The process is initiated when the user requests a server-side protected resource, such as a web page within the domain "ibm.com" (step 152). The terms "server-side" and "client-side" refer to actions or entities at a server or a client, respectively, within a networked environment. The web browser (or 10 associated application or applet) generates an HTTP request (step 153) that is sent to the web server that is hosting the domain "ibm.com". The terms "request" and "response" should be understood to comprise data formatting that is appropriate for the transfer of information that is involved in 15 a particular operation, such as messages, communication protocol information, or other associated information.

The server determines that it does not have an active session for the client (step 154), so the server initiates and completes the establishment of an SSL (Secure Sockets Layer) 20 session between the server and the client (step 155), which entails multiple transfers of information between the client and the server. After an SSL session is established, subsequent communication messages are transferred within the SSL session; any secret information remains secure because 25 of the encrypted communications within the SSL session.

However, the server needs to determine the identity of the user before allowing the user to have access to protected resources, so the server requires the user to perform an authentication process by sending the client some type of 30 authentication challenge (step 156). The authentication challenge may be in various formats, such as an HTML form. The user then provides the requested or required information (step 157), such as a username or other type of user identifier along with an associated password or other form of secret information.

35 The authentication response information is sent to the server (step 158), at which point the server authenticates the user or client (step 159), e.g., by retrieving previously submitted registration information and matching the presented authentication information with the user's stored information. Assuming the authentication is successful, an active session is established for the authenticated user or client. The server creates a session identifier for the client, and any subsequent request messages from the client within the session 40 would be accompanied by the session identifier.

The server then retrieves the originally requested web page and sends an HTTP response message to the client (step 160), thereby fulfilling the user's original request for the protected resource. At that point, the user may request another page within "ibm.com" (step 161) by clicking a hypertext link within a browser window, and the browser sends another HTTP request message to the server (step 162). At that point, the server recognizes that the user has an active session (step 50 163) because the user's session identifier is returned to the server in the HTTP request message, and the server sends the requested web page back to the client in another HTTP response message (step 164). Although FIG. 1C depicts a typical prior art process, it should be noted that other alternative session state management techniques may be depicted, such as URL rewriting or using cookies to identify users with active sessions, which may include using the same cookie that is used to provide proof of authentication.

With reference now to FIG. 1D, a diagram illustrates a 55 typical Web-based environment in which the present invention may be implemented. In this environment, a user of browser 170 at client 171 desires to access a protected

US 7,631,346 B2

7

resource on web application server 172 in DNS domain 173, or on web application server 174 in DNS domain 175.

In a manner similar to that shown in FIG. 1C, a user can request a protected resource at one of many domains. In contrast to FIG. 1C, which shows only a single server at a particular domain, each domain in FIG. 1D has multiple servers. In particular, each domain may have an associated authentication server 176 and 177.

In this example, after client 171 issues a request for a protected resource at domain 173, web application server 172 determines that it does not have an active session for client 171, and it requests that authentication server 176 perform an appropriate authentication operation with client 171. Authentication server 176 communicates the result of the authentication operation to web application server 172. If the user (or browser 170 or client 171 on behalf of the user) is successfully authenticated, then web application server 172 establishes a session for client 171 and returns the requested protected resource. Typically, once the user is authenticated by the authentication server, a cookie may be set and stored in a cookie cache in the browser. FIG. 1D is merely an example of one manner in which the processing resources of a domain may be shared amongst multiple servers, particularly to perform authentication operations.

In a similar manner, after client 171 issues a request for a protected resource at domain 175, authentication server 177 performs an appropriate authentication operation with client 171, after which web application server 174 establishes a session for client 171 and returns the requested protected resource. Hence, FIG. 1D illustrates that client 171 may have multiple concurrent sessions in different domains yet is required to complete multiple authentication operations to establish those concurrent sessions.

With reference now to FIG. 1E, a block diagram depicts an example of a typical online transaction that might require multiple authentication operations from a user. Referring again to FIG. 1C and FIG. 1D, a user may be required to complete an authentication operation prior to gaining access to a controlled resource, as shown in FIG. 1C. Although not shown in FIG. 1C, an authentication manager may be deployed on server 151 to retrieve and employ user information that is required to authenticate a user. As shown in FIG. 1D, a user may have multiple current sessions within different domains 173 and 175, and although they are not shown in FIG. 1D, each domain may employ an authentication manager in place of or in addition to the authentication servers. In a similar manner, FIG. 1E also depicts a set of domains, each of which support some type of authentication manager. FIG. 1E illustrates some of the difficulties that a user may experience when accessing multiple domains that require the user to complete an authentication operation for each domain.

User 190 may be registered at ISP domain 191, which may support authentication manager 192 that authenticates user 190 for the purpose of completing transactions with respect to domain 191. ISP domain 191 may be an Internet Service Provider (ISP) that provides Internet connection services, email services, and possibly other e-commerce services. Alternatively, ISP domain 191 may be an Internet portal that is frequently accessed by user 190.

Similarly, domains 193, 195, and 197 represent typical web service providers. Government domain 193 supports authentication manager 194 that authenticates users for completing various government-related transactions. Banking domain 195 supports authentication manager 196 that authenticates users for completing transactions with an online bank. E-commerce domain 197 supports authentication manager 198 that authenticates users for completing online purchases.

8

As noted previously, when a user attempts to move from one domain to another domain within the Internet or World Wide Web by accessing resources at the different domains, a user may be subjected to multiple user authentication requests or requirements, which can significantly slow the user's progress across a set of domains. Using FIG. 1E as an exemplary environment, user 190 may be involved in a complicated online transaction with e-commerce domain 197 in which the user is attempting to purchase an on-line service that is limited to users who are at least 18 years old and who have a valid driver license, a valid credit card, and a U.S. bank account. This online transaction may involve domains 191, 193, 195, and 197.

Typically, a user might not maintain an identity and/or attributes within each domain that participates in a typical online transaction. In this example, user 190 may have registered his or her identity with the user's ISP, but to complete the online transaction, the user might also be required to authenticate to domains 193, 195, and 197. If each of the domains does not maintain an identity for the user, then the user's online transaction may fail. Even if the user can be authenticated by each domain, it is not guaranteed that the different domains can transfer information between themselves in order to complete the user's transaction.

Given the preceding brief description of some current technology, the description of the remaining figures relates to federated computer environments in which the present invention may operate. Prior to discussing the present invention in more detail, however, some terminology is introduced.

Terminology

The terms "entity" or "party" generally refers to an organization, an individual, or a system that operates on behalf of an organization, an individual, or another system. The term "domain" connotes additional characteristics within a network environment, but the terms "entity", "party", and "domain" can be used interchangeably. For example, the term "domain" may also refer to a DNS (Domain Name System) domain, or more generally, to a data processing system that includes various devices and applications that appear as a logical unit to exterior entities.

The terms "request" and "response" should be understood to comprise data formatting that is appropriate for the transfer of information that is involved in a particular operation, such as messages, communication protocol information, or other associated information. A protected resource is a resource (an application, an object, a document, a page, a file, executable code, or other computational resource, communication-type resource, etc.) for which access is controlled or restricted.

A token provides direct evidence of a successful operation and is produced by the entity that performs the operation, e.g., an authentication token that is generated after a successful authentication operation. A Kerberos token is one example of an authentication token that may be used with the present invention. More information on Kerberos may be found in Kohl et al., "The Kerberos Network Authentication Service (V5)", Internet Engineering Task Force (IETF) Request for Comments (RFC) 1510, 09/1993.

An assertion provides indirect evidence of some action. Assertions may provide indirect evidence of identity, authentication, attributes, authorization decisions, or other information and/or operations. An authentication assertion provides indirect evidence of authentication by an entity that is not the authentication service but that listened to the authentication service.

A Security Assertion Markup Language (SAML) assertion is an example of a possible assertion format that may be used with the present invention. SAML has been promulgated by

US 7,631,346 B2

9

the Organization for the Advancement of Structured Information Standards (OASIS), which is a non-profit, global consortium. SAML is described in “Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML)”, Committee Specification 01, May 31, 2002, as follows:

The Security Assertion Markup Language (SAML) is an XML-based framework for exchanging security information. This security information is expressed in the form of assertions about subjects, where a subject is an entity (either human or computer) that has an identity in some security domain. A typical example of a subject is a person, identified by his or her email address in a particular Internet DNS domain. Assertions can convey information about authentication acts performed by subjects, attributes of subjects, and authorization decisions about whether subjects are allowed to access certain resources. Assertions are represented as XML constructs and have a nested structure, whereby a single assertion might contain several different internal statements about authentication, authorization, and attributes. Note that assertions containing authentication statements merely describe acts of authentication that happened previously. Assertions are issued by SAML authorities, namely, authentication authorities, attribute authorities, and policy decision points. SAML defines a protocol by which clients can request assertions from SAML authorities and get a response from them. This protocol, consisting of XML-based request and response message formats, can be bound to many different underlying communications and transport protocols; SAML currently defines one binding, to SOAP over HTTP. SAML authorities can use various sources of information, such as external policy stores and assertions that were received as input in requests, in creating their responses. Thus, while clients always consume assertions, SAML authorities can be both producers and consumers of assertions.

The SAML specification states that an assertion is a package of information that supplies one or more statements made by an issuer. SAML allows issuers to make three different kinds of assertion statements: authentication, in which the specified subject was authenticated by a particular means at a particular time; authorization, in which a request to allow the specified subject to access the specified resource has been granted or denied; and attribute, in which the specified subject is associated with the supplied attributes. As discussed further below, various assertion formats can be translated to other assertion formats when necessary.

Authentication is the process of validating a set of credentials that are provided by a user or on behalf of a user. Authentication is accomplished by verifying something that a user knows, something that a user has, or something that the user is, i.e. some physical characteristic about the user. Something that a user knows may include a shared secret, such as a user’s password, or by verifying something that is known only to a particular user, such as a user’s cryptographic key. Something that a user has may include a smartcard or hardware token. Some physical characteristic about the user might include a biometric input, such as a fingerprint or a retinal map.

An authentication credential is a set of challenge/response information that is used in various authentication protocols. For example, a username and password combination is the most familiar form of authentication credentials. Other forms of authentication credential may include various forms of challenge/response information, Public Key Infrastructure (PKI) certificates, smartcards, biometrics, etc. An authentica-

10

tion credential is differentiated from an authentication assertion: an authentication credential is presented by a user as part of an authentication protocol sequence with an authentication server or service, and an authentication assertion is a statement about the successful presentation and validation of a user’s authentication credentials, subsequently transferred between entities when necessary.

Federation Model for Computing Environment that May Incorporate the Present Invention

In the context of the World Wide Web, users are coming to expect the ability to jump from interacting with an application on one Internet domain to another application on another domain with minimal regard to the information barriers between each particular domain. Users do not want the frustration that is caused by having to authenticate to multiple domains for a single transaction. In other words, users expect that organizations should interoperate, but users generally want domains to respect their privacy. In addition, users may prefer to limit the domains that permanently store private information. These user expectations exist in a rapidly evolving heterogeneous environment in which many enterprises and organizations are promulgating competing authentication techniques.

The present invention is supported within a federation model that allows enterprises to provide a single-sign-on experience to a user. In other words, the present invention may be implemented within a federated, heterogeneous environment. As an example of a transaction that would benefit from a federated, heterogeneous environment, referring again to FIG. 1E, user 190 is able to authenticate to domain 191 and then have domain 191 provide the appropriate assertions to each downstream domain that might be involved in a transaction. These downstream domains need to be able to understand and trust authentication assertions and/or other types of assertions, even though there are no pre-established assertion formats between domain 191 and these other downstream domains. In addition to recognizing the assertions, the downstream domains need to be able to translate the identity contained within an assertion to an identity that represents user 190 within a particular domain, even though there is no pre-established identity mapping relationship. It should be noted, though, that the present invention is applicable to various types of domains and is not limited to ISP-type domains that are represented within FIG. 1E as exemplary domains.

The present invention is supported within a federated environment. In general, an enterprise has its own user registry and maintains relationships with its own set of users. Each enterprise typically has its own means of authenticating these users. However, the federated scheme for use with the present invention allows enterprises to cooperate in a collective manner such that users in one enterprise can leverage relationships with a set of enterprises through an enterprise’s participation in a federation of enterprises. Users can be granted access to resources at any of the federated enterprises as if they had a direct relationship with each enterprise. Users are not required to register at each business of interest, and users are not constantly required to identify and authenticate themselves. Hence, within this federated environment, an authentication scheme allows for a single-sign-on experience within the rapidly evolving heterogeneous environments in information technology.

In the context of the present invention, a federation is a set of distinct entities, such as enterprises, organizations, institutions, etc., that cooperate to provide a single-sign-on, ease-of-use experience to a user; a federated environment differs from a typical single-sign-on environment in that two enterprises need not have a direct, pre-established, relationship

US 7,631,346 B2

11

defining how and what information to transfer about a user. Within a federated environment, entities provide services which deal with authenticating users, accepting authentication assertions, e.g., authentication tokens, that are presented by other entities, and providing some form of translation of the identity of the vouched-for user into one that is understood within the local entity.

Federation eases the administrative burden on service providers. A service provider can rely on its trust relationships with respect to the federation as a whole; the service provider does not need to manage authentication information, such as user password information, because it can rely on authentication that is accomplished by a user's authentication home domain or an identity provider.

The system that supports the present invention also concerns a federated identity management system that establishes a foundation in which loosely coupled authentication, user enrollment, user profile management and/or authorization services collaborate across security domains. Federated identity management allows services residing in disparate security domains to securely interoperate and collaborate even though there may be differences in the underlying security mechanisms and operating system platforms at these disparate domains.

Identity Provider vs. Service Provider

As mentioned above and as explained in more detail further below, a federated environment provides significant user benefits. A federated environment allows a user to authenticate at a first entity, which may act as an issuing party to issue an authentication assertion about the user for use at a second entity. The user can then access protected resources at a second, distinct entity, termed the relying party, by presenting the authentication assertion that was issued by the first entity without having to explicitly re-authenticate at the second entity. Information that is passed from an issuing party to a relying party is in the form of an assertion, and this assertion may contain different types of information in the form of statements. For example, an assertion may be a statement about the authenticated identity of a user, or it may be a statement about user attribute information that is associated with a particular user.

With reference now to FIG. 2, a block diagram depicts the terminology of the federated environment with respect to a transaction that is initiated by a user to a first federated enterprise, which, in response, invokes actions at downstream entities within the federated environment. FIG. 2 shows that the terminology may differ depending on the perspective of an entity within the federation for a given federated operation. More specifically, FIG. 2 illustrates that a computing environment that supports the present invention supports the transitivity of trust and the transitivity of the authentication assertion process; a domain or an entity can issue an assertion based on its trust in an identity as asserted by another domain or another entity.

User 202 initiates a transaction through a request for a protected resource at enterprise 204. If user 202 has been authenticated by enterprise 204 or will eventually be authenticated by enterprise 204 during the course of a transaction, then enterprise 204 may be termed the user's home domain for this federated session. Assuming that the transaction requires some type of operation by enterprise 206 and enterprise 204 transfers an assertion to enterprise 206, then enterprise 204 is the issuing entity with respect to the particular operation, and enterprise 206 is the relying entity for the operation.

The issuing entity issues an assertion for use by the relying domain; an issuing entity is usually, but not necessarily, the

12

user's home domain or the user's identity provider. Hence, it would usually be the case that the issuing party has authenticated the user using a typical authentication operation. However, it is possible that the issuing party has previously acted as a relying party whereby it received an assertion from a different issuing party. In other words, since a user-initiated transaction may cascade through a series of enterprises within a federated environment, a receiving party may subsequently act as an issuing party for a downstream transaction. In general, any entity that has the ability to issue authentication assertions on behalf of a user can act as an issuing entity.

The relying entity is an entity that receives an assertion from an issuing entity. The relying party is able to accept, trust, and understand an assertion that is issued by a third party on behalf of the user, i.e. the issuing entity; it is generally the relying entity's duty to use an appropriate authentication authority to interpret an authentication assertion. A relying party is an entity that relies on an assertion that is presented on behalf of a user or another entity. In this manner, a user can be given a single-sign-on experience at the relying entity instead of requiring the relying entity to prompt the user for the user's authentication credentials as part of an interactive session with the user.

Referring again to FIG. 2, assuming that the transaction requires further operations such that enterprise 206 transfers an assertion to enterprise 208, then enterprise 206 is an upstream entity that acts as the issuing entity with respect to the subsequent or secondary transaction operation, and enterprise 208 is a downstream entity that acts as the relying entity for the operation; in this case, enterprise 208 may be regarded as another downstream entity with respect to the original transaction, although the subsequent transaction can also be described with respect to only two entities.

As shown in FIG. 2, a federated entity may act as a user's home domain, which provides identity information and attribute information about federated users. An entity within a federated computing environment that provides identity information, identity or authentication assertions, or identity services may be termed an identity provider. Other entities or federation partners within the same federation may rely on an identity provider for primary management of a user's authentication credentials, e.g., accepting a single-sign-on token that is provided by the user's identity provider; a domain at which the user authenticates may be termed the user's (authentication) home domain. The identity provider may be physically supported by the user's employer, the user's ISP, or some other commercial entity.

An identity provider is a specific type of service that provides identity information as a service to other entities within a federated computing environment. With respect to most federated transactions, an issuing party for an authentication assertion would usually be an identity provider; any other entity can be distinguished from the identity provider. Any other entity that provides a service within the federated computing environment can be categorized as a service provider. Once a user has authenticated to the identity provider, other entities or enterprises in the federation may be regarded as merely service providers for the duration of a given federated session or a given federated transaction.

In some circumstances, there may be multiple entities within a federated environment that may act as identity providers for a user. For example, the user may have accounts at multiple federated domains, each of which is able to act as an identity provider for the user; these domains do not necessarily have information about the other domains nor about a user's identity at a different domain.

US 7,631,346 B2

13

Although it may be possible that there could be multiple enterprises within a federated environment that may act as identity providers, e.g., because there may be multiple enterprises that have the ability to generate and validate a user's authentication credentials, etc., a federated transaction usually involves only a single identity provider. If there is only a single federated entity that is able to authenticate a user, e.g., because there is one and only one entity within the federation with which the user has performed a federated enrollment or registration operation, then it would be expected that this entity would act as the user's identity provider in order to support the user's transactions throughout the federated environment.

Within some federated transactions that require the inter-operation of multiple service providers, a downstream service provider may accept an assertion from an upstream service provider; the conditions in which an upstream service provider may act as an issuing entity to a downstream service provider that is acting as a relying party may depend upon the type of trust relationship between the service providers and the type of transaction between the service providers. Within the scope of a simple federated transaction, however, there is only one entity that acts as an issuing entity.

The present invention may be supported within a given computing environment in which a federated infrastructure can be added to existing systems while minimizing the impact on an existing, non-federated architecture. Hence, operations, including authentication operations, at any given enterprise or service provider are not necessarily altered by the fact that an entity may also participate within a federated environment. In other words, even though an entity's computing systems may be integrated into a federated environment, a user may be able to continue to perform various operations, including authentication operations, directly with an enterprise in a non-federated manner. However, the user may be able to have the same end-user experience while performing a federated operation with respect to a given entity as if the user had performed a similar operation with the given entity in a non-federated manner. Hence, it should be noted that not all of a given enterprise's users necessarily participate federated transactions when the given enterprise participates in a federation; some of the enterprise's users may interact with the enterprise's computing systems without performing any federated transactions.

Moreover, user registration within the computing environment of a given enterprise, e.g., establishment of a user account in a computer system, is not necessarily altered by the fact that the enterprise may also participate within a federated environment. For example, a user may still establish an account at a domain through a legacy or pre-existing registration process that is independent of a federated environment. Hence, in some cases, the establishment of a user account at an enterprise may or may not include the establishment of account information that is valid across a federation when the enterprise participates within a federated computing environment.

Federated Architecture—Federation Front-End for Legacy Systems

With reference now to FIG. 3, a block diagram depicts the integration of pre-existing data processing systems at a given domain with some federated architecture components that may be used to support an embodiment of the present invention. A federated environment includes federated entities that provide a variety of services for users. User 312 interacts with client device 314, which may support browser application 316 and various other client applications 318. User 312 is distinct from client device 314, browser 316, or any other

14

software that acts as interface between user and other devices and services. In some cases, the following description may make a distinction between the user acting explicitly within a client application and a client application that is acting on behalf of the user. In general, though, a requester is an intermediary, such as a client-based application, browser, SOAP client, etc., that may be assumed to act on behalf of the user.

Browser application 316 may be a typical browser, including those found on mobile devices, that comprises many modules, such as HTTP communication component 320 and markup language (ML) interpreter 322. Browser application 316 may also support plug-ins, such as web services client 324, and/or downloadable applets, which may or may not require a virtual machine runtime environment. Web services client 324 may use Simple Object Access Protocol (SOAP), which is a lightweight protocol for defining the exchange of structured and typed information in a decentralized, distributed environment. SOAP is an XML-based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it; a set of encoding rules for expressing instances of application-defined datatypes; and a convention for representing remote procedure calls and responses. User 312 may access web-based services using browser application 316, but user 312 may also access web services through other web service clients on client device 314. Some of the federated operations may employ HTTP redirection via the user's browser to exchange information between entities in a federated environment. However, it should be noted that the present invention may be supported over a variety of communication protocols and is not meant to be limited to HTTP-based communications. For example, the entities in the federated environment may communicate directly when necessary; messages are not required to be redirected through the user's browser.

The present invention may be supported in a manner such that components that are required for a federated environment can be integrated with pre-existing systems. FIG. 3 depicts one embodiment for implementing these components as a front-end to a pre-existing system. The pre-existing components at a federated domain can be considered as legacy applications or back-end processing components 330, which include authentication service runtime (ASR) servers 332 in a manner similar to that shown in FIG. 4. ASR servers 332 are responsible for authenticating users when the domain controls access to application servers 334, which can be considered to generate, retrieve, or otherwise support or process protected resources 335. The domain may continue to use legacy user registration application 336 to register users for access to application servers 334. Information that is needed to authenticate a registered user with respect to legacy operations is stored in enterprise user registry 338; enterprise user registry 338 may be accessible to federation components as well.

After joining a federated environment, the domain may continue to operate without the intervention of federated components. In other words, the domain may be configured so that users may continue to access particular application servers or other protected resources directly without going through a point-of-contact server or other component implementing this point-of-contact server functionality; a user that accesses a system in this manner would experience typical authentication flows and typical access. In doing so, however, a user that directly accesses the legacy system would not be able to establish a federated session that is known to the domain's point-of-contact server.

US 7,631,346 B2

15

The domain's legacy functionality can be integrated into a federated environment through the use of federation front-end processing **340**, which includes point-of-contact server **342** and trust proxy server **344** (or more simply, trust proxy **344** or trust service **344**) which itself interacts with Security Token Service (STS) **346**, which are described in more detail below with respect to FIG. 4. Federation configuration application **348** allows an administrative user to configure the federation front-end components to allow them to interface with the legacy back-end components through federation interface unit **350**. Federated functionality may be implemented in distinct system components or modules. In a preferred embodiment, most of the functionality for performing federation operations may be implemented by a collection of logical components within a single federation application; federated user lifecycle management application **352** includes trust service **344** along with single-sign-on protocol service (SPS) **354**. Trust service **344** may comprise identity-and-attribute service (I&AS) **356**, which is responsible for identity mapping operations, attribute retrieval, etc., as part of federation functionality. Identity-and-attribute service **356** may also be employed by single-sign-on protocol service **354** during single-sign-on operations. A federation user registry **358** may be employed in certain circumstances to maintain user-related information for federation-specific purposes.

Legacy or pre-existing authentication services at a given enterprise may use various, well known, authentication methods or tokens, such as username/password or smart card token-based information. However, in a preferred federated computing system for supporting the present invention, the functionality of a legacy authentication service can be used in a federated environment through the use of point-of-contact servers. Users may continue to access a legacy authentication server directly without going through a point-of-contact server, although a user that accesses a system in this manner would experience typical authentication flows and typical access; a user that directly accesses a legacy authentication system would not be able to generate a federated authentication assertion as proof of identity in accordance with the present invention. One of the roles of the federation front-end is to translate a federated authentication token received at a point-of-contact server into a format understood by a legacy authentication service. Hence, a user accessing the federated environment via the point-of-contact server would not necessarily be required to re-authenticate to the legacy authentication service. Preferably, the user would be authenticated to a legacy authentication service by a combination of the point-of-contact server and a trust proxy such that it appears as if the user was engaged in an authentication dialog.

Federated Architecture—Point-of-Contact Servers, Trust Proxies, and Trust Brokers

With reference now to FIG. 4, a block diagram depicts an example of a manner in which some components within a federated architecture may be used to establish trust relationships to support an implementation of the present invention. A federated environment includes federated enterprises or similar entities that provide a variety of services for users. A user, through an application on a client device, may attempt to access resources at various entities, such as enterprise **410**. A point-of-contact server at each federated enterprise, such as point-of-contact (POC) server **412** at enterprise **410**, is the entry point into the federated environment for requests from a client to access resources that are supported and made available by enterprise **410**. The point-of-contact server minimizes the impact on existing components within an existing, non-federated architecture, e.g., legacy systems, because the point-of-contact server handles many of the federation

16

requirements. The point-of-contact server provides session management, protocol conversion, and possibly initiates authentication and/or attribute assertion conversion. For example, the point-of-contact server may translate HTTP or HTTPS messages to SOAP and vice versa. As explained in more detail further below, the point-of-contact server may also be used to invoke a trust proxy to translate assertions, e.g., a SAML token received from an issuing party can be translated into a Kerberos token understood by a receiving party.

A trust service (also termed a trust proxy, a trust proxy server, or a trust service), such as trust proxy (TP) **414** at enterprise **410**, establishes and maintains a trust relationship between two entities in a federation. A trust service generally has the ability to handle authentication token format translation (through the security token service, which is described in more detail further below) from a format used by the issuing party to one understood by the receiving party.

Together, the use of a point-of-contact server and a trust service minimize the impact of implementing a federated architecture on an existing, non-federated set of systems. Hence, the exemplary federated architecture requires the implementation of at least one point-of-contact server and at least one trust service per federated entity, whether the entity is an enterprise, a domain, or other logical or physical entity. The exemplary federated architecture, though, does not necessarily require any changes to the existing, non-federated set of systems. Preferably, there is a single trust service for a given federated entity, although there may be multiple instances of a trust service component for availability purposes, or there may be multiple trust services for a variety of smaller entities within a federated entity, e.g., separate subsidiaries within an enterprise. It is possible that a given entity could belong to more than one federation, although this scenario would not necessarily require multiple trust services as a single trust service may be able to manage trust relationships within multiple federations.

One role of a trust service may be to determine or to be responsible for determining the required token type by another domain and/or the trust service in that domain. A trust service has the ability or the responsibility to handle authentication token format translation from a format used by the issuing party to one understood by the receiving party. Trust service **414** may also be responsible for any user identity translation or attribute translation that occurs for enterprise **410**, or this responsibility may be supported by a distinct identity-and-attribute service, e.g., such as identity-and-attribute service **356** as shown in FIG. 3. In addition, a trust service can support the implementation of aliases as representatives of a user identity that uniquely identify a user without providing any additional information about the user's real world identity. Furthermore, a trust proxy can issue authorization and/or session credentials for use by the point-of-contact server. However, a trust service may invoke a trust broker for assistance, as described further below. Identity translation may be required to map a user's identity and attributes as known to an issuing party to one that is meaningful to a receiving party. This translation may be invoked by either a trust service at an issuing entity, a trust service at a receiving entity, or both.

Trust service **414**, or a distinct identity-and-attribute service as mentioned above, may include (or interact with) an internalized component, shown as security token service (STS) component **416**, which will provide token translation and will invoke authentication service runtime (ASR) **418** to validate and generate tokens. The security token service provides the token issuance and validation services required by

US 7,631,346 B2

17

the trust service, which may include identity translation. The security token service therefore includes an interface to existing authentication service runtimes, or it incorporates authentication service runtimes into the service itself. Rather than being internalized within the trust service, the security token service component may also be implemented as a stand-alone component, e.g., to be invoked by the trust service, or it may be internalized within a transaction server, e.g., as part of an application server.

For example, an security token service component may receive a request to issue a Kerberos token. As part of the authentication information of the user for whom the token is to be created, the request may contain a binary token containing a username and password. The security token service component will validate the username and password against, e.g., an LDAP runtime (typical authentication) and will invoke a Kerberos KDC (Key Distribution Center) to generate a Kerberos ticket for this user. This token is returned to the trust service for use within the enterprise; however, this use may include externalizing the token for transfer to another domain in the federation.

In a manner similar to that described with respect to FIG. 1D, a user may desire to access resources at multiple enterprises within a federated environment, such as both enterprise 410 and enterprise 420. In a manner similar to that described above for enterprise 410, enterprise 420 comprises point-of-contact server 422, trust service 424, security token service (STS) 426, and authentication service runtime 428. Although the user may directly initiate separate transactions with each enterprise, the user may initiate a transaction with enterprise 410 which cascades throughout the federated environment. Enterprise 410 may require collaboration with multiple other enterprises within the federated environment, such as enterprise 420, to complete a particular transaction, even though the user may not have been aware of this necessity when the user initiated a transaction. Enterprise 420 becomes involved as a downstream entity, and enterprise 410 may present a assertion to enterprise 420 if necessary in order to further the user's federated transaction.

It may be the case that a trust service does not know how to interpret the authentication token that is received by an associated point-of-contact server and/or how to translate a given user identity and attributes. In this case, the trust service may choose to invoke functionality at a trust broker component, such as trust broker 430. A trust broker maintains relationships with individual trust proxies/services, thereby providing transitive trust between trust services. Using a trust broker allows each entity within a federated environment, such enterprises 410 and 420, to establish a trust relationship with the trust broker rather than establishing multiple individual trust relationships with each entity in the federated environment. For example, when enterprise 420 becomes involved as a downstream entity for a transaction initiated by a user at enterprise 410, trust service 414 at enterprise 410 can be assured that trust service 424 at enterprise 420 can understand an assertion from trust service 414 by invoking assistance at trust broker 430 if necessary. Although FIG. 4 depicts the federated environment with a single trust broker, a federated environment may have multiple trust brokers.

It should be noted that although FIG. 4 depicts point-of-contact server 412, trust service 414, security token service component 416, and authentication service runtime 418 as distinct entities, it is not necessary for these components to be implemented on separate components. For example, it is possible for the functionality of these separate components to be implemented as a single application, as applications on a single physical device, or as distributed applications on mul-

18

iple physical devices. In addition, FIG. 4 depicts a single point-of-contact server, a single trust service, and a single security token server for an enterprise, but an alternative configuration may include multiple point-of-contact servers, multiple trust services, and multiple security token servers for each enterprise. The point-of-contact server, the trust service, the security token service, and other federated entities may be implemented in various forms, such as software applications, objects, modules, software libraries, etc.

A trust service/STS may be capable of accepting and validating many different authentication credentials, including traditional credentials such as a username and password combinations and Kerberos tickets, and federated authentication token formats, including authentication tokens produced by a third party. A trust service/STS may allow the acceptance of an authentication token as proof of authentication elsewhere. The authentication token is produced by an issuing party and is used to indicate that a user has already authenticated to that issuing party. The issuing party produces the authentication token as a means of asserting the authenticated identity of a user. A trust service/STS is also able to process attribute tokens or tokens that are used to secure communication sessions or conversations, e.g., those that are used to manage session information in a manner similar to an SSL session identifier.

A security token service invokes an authentication service runtime as necessary. The authentication service runtime supports an authentication service capable of authenticating a user. The authentication service acts as an authentication authority that provides indications of successful or failed authentication attempts via authentication responses. The trust service/STS may internalize an authentication service, e.g., a scenario in which there is a brand-new installation of a web service that does not need to interact with an existing legacy infrastructure. Otherwise, the security token service component will invoke external authentication services for validation of authentication tokens. For example, the security token service component could "unpack" a binary token containing a username/password and then use an LDAP service to access a user registry to validate the presented credentials.

When used by another component such as an application server, the security token service component can be used to produce tokens required for single-sign-on to legacy authentication systems; this functionality may be combined with or replaced by functionality within a single-sign-on protocol service, such as SPS 354 that is shown in FIG. 3. Hence, the security token service component can be used for token translation for internal purposes, i.e. within an enterprise, and for external purposes, i.e. across enterprises in a federation. As an example of an internal purpose, a Web application server may interface to a mainframe via an IBM CICS (Customer Information Control System) transaction gateway; CICS is a family of application servers and connectors that provides enterprise-level online transaction management and connectivity for mission-critical applications. The Web application server may invoke the security token service component to translate a Kerberos ticket (as used internally by the Web application server) to an IBM RACF® passticket required by the CICS transaction gateway.

The entities that are shown in FIG. 4 can be explained using the terminology that was introduced above, e.g., "identity provider" and "service provider". As part of establishing and maintaining trust relationships, an identity provider's trust service can determine what token types are required/accepted by a service provider's trust service. Thus, trust services use this information when invoking token services from a security token service. When an identity provider's trust service is

US 7,631,346 B2

19

required to produce an authentication assertion for a service provider, the trust service determines the required token type and requests the appropriate token from the security token service.

When a service provider's trust service receives an authentication assertion from an identity provider, the trust service knows what type of assertion that it expected and what type of assertion that it needs for internal use within the service provider. The service provider's trust service therefore requests that the security token service generate the required internal-use token based on the token in the received authentication assertion.

Both trust services and trust brokers have the ability to translate an assertion received from an identity provider into a format that is understood by a service provider. The trust broker has the ability to interpret the assertion format (or formats) for each of the trust services with whom there is a direct trust relationship, thereby allowing the trust broker to provide assertion translation between an identity provider and a service provider. This translation can be requested by either party through its local trust service. Thus, the identity provider's trust service can request translation of an assertion before it is sent to the service provider. Likewise, the service provider's trust service can request translation of an assertion received from an identity provider.

Assertion translation comprises user identity translation, authentication assertion translation, attribute assertion translation, or other forms of assertion translation. Reiterating the point above, assertion translation is handled by the trust components within a federation, e.g., trust services and trust brokers. A trust service may perform the translation locally, either at the identity provider or at the service provider, or a trust service may invoke assistance from a trust broker.

Assuming that an identity provider and a service provider already have individual trust relationships with a trust broker, the trust broker can dynamically create, i.e. broker, new trust relationships between issuing parties and relying parties if necessary. After the initial trust relationship brokering operation that is provided by the trust broker, the identity provider and the service provider may directly maintain the relationship so that the trust broker need not be invoked for future translation requirements. It should be noted that translation of authentication tokens can happen at three possible places: the identity provider's trust service, the service provider's trust service, and the trust broker. Preferably, the identity provider's trust service generates an authentication assertion that is understood by the trust broker to send to the service provider. The service provider then requests a translation of this token from the trust broker into a format recognizable by the service provider. Token translation may occur before transmission, after transmission, or both before and after transmission of the authentication assertion.

Trust Relationships within Federated Architecture

Within an exemplary federated environment that is able to support the present invention, there are two types of "trust domains" that must be managed: enterprise trust domains and federation trust domains. The differences between these two types of trust domain are based in part on the business agreements governing the trust relationships with the trust domain and the technology used to establish trust. An enterprise trust domain contains those components that are managed by the enterprise; all components within that trust domain may implicitly trust each other. In general, there are no business agreements required to establish trust within an enterprise because the deployed technology creates inherent trust within an enterprise, e.g., by requiring mutually authenticated SSL sessions between components or by placing components

20

within a single, tightly controlled data center such that physical control and proximity demonstrate implicit trust. Referring to FIG. 2B, the legacy applications and back-end processing systems may represent an enterprise trust domain, wherein the components communicate on a secure internal network.

Federation trust domains are those that cross enterprise boundaries; from one perspective, a federation trust domain may represent trust relationships between distinct enterprise trust domains. Federation trust domains are established through trust proxies across enterprise boundaries between federation partners. Trust relationships involve some sort of a bootstrapping process by which initial trust is established between trust proxies. Part of this bootstrap process may include the establishment of shared secret keys and rules that define the expected and/or allowed token types and identifier translations. In general, this bootstrapping process can be implemented out-of-band as this process may also include the establishment of business agreements that govern an enterprise's participation in a federation and the liabilities associated with this participation.

There are a number of possible mechanisms for establishing trust in a federated business model. In a federation model, a fundamental notion of trust between the federation participants is required for business reasons in order to provide a level of assurance that the assertions (including tokens and attribute information) that are transferred between the participants are valid. If there is no trust relationship, then the service provider cannot depend upon the assertions received from the identity provider; they cannot be used by the service provider to determine how to interpret any information received from the identity provider.

For example, a large corporation may want to link several thousand global customers, and the corporation could use non-federated solutions. As a first example, the corporation could require global customers to use a digital certificate from a commercial certificate authority to establish mutual trust. The commercial certificate authority enables the servers at the corporation to trust servers located at each of the global customers. As a second example, the corporation could implement third-party trust using Kerberos; the corporation and its global customers could implement a trusted third-party Kerberos domain service that implements shared-secret-based trust. As a third example, the corporation could establish a private scheme with a proprietary security message token that is mutually trusted by the servers of its global customers.

Any one of these approaches may be acceptable if the corporation needed to manage trust relationships with a small number of global customers, but this may become unmanageable if there are hundreds or thousands of potential federation partners. For example, while it may be possible for the corporation to force its smaller partners to implement a private scheme, it is unlikely that the corporation will be able to impose many requirements on its larger partners.

An enterprise may employ trust relationships established and maintained through trust proxies and possibly trust brokers. An advantage of the exemplary federated architecture that is shown in the figures is that it does not impose additional requirements above and beyond the current infrastructures of an enterprise and its potential federation partners.

However, this exemplary federation architecture does not relieve an enterprise and its potential federation partners from the preliminary work required to establish business and liability agreements that are required for participation in the federation. In addition, the participants cannot ignore the technological bootstrapping of a trust relationship. The

US 7,631,346 B2

21

exemplary federation architecture allows this bootstrapping to be flexible, e.g., a first federation partner can issue a Kerberos ticket with certain information, while a second federation partner can issue a SAML authentication assertion with certain information.

In the exemplary federation architecture, the trust relationships are managed by the trust proxies, which may include (or may interact with) a security token service that validates and translates a token that is received from an identity provider based on the pre-established relationship between two trust proxies. In situations where it is not feasible for a federated enterprise to establish trust relationships (and token translation) with another federated enterprise, a trust broker may be invoked; however, the federated enterprise would need to establish a relationship with a trust broker.

With reference now to FIG. 5, a block diagram depicts an exemplary set of trust relationships between federated domains using trust proxies and a trust broker in accordance with an exemplary federated architecture that is able to support the present invention. Although FIG. 4 introduced the trust broker, FIG. 5 illustrates the importance of transitive trust relationships within the exemplary federated architecture.

Federated domains **502-506** incorporate trust proxies **508-512**, respectively. Trust proxy **508** has direct trust relationship **514** with trust proxy **510**. Trust broker **520** has direct trust relationship **516** with trust proxy **510**, and trust broker **520** has direct trust relationship **518** with trust proxy **512**. Trust broker **520** is used to establish, on behalf of a federation participant, a trust relationship based on transitive trust with other federation partners. The principle of transitive trust allows trust proxy **510** and trust proxy **512** to have brokered trust relationship **522** via trust broker **520**. Neither trust proxy **510** nor **512** need to know how to translate or validate the other's assertions; the trust broker may be invoked to translate an assertion into one that is valid, trusted, and understood at the other trust proxy.

Business agreements that specify contractual obligations and liabilities with respect to the trust relationships between federated enterprises can be expressed in XML through the use of the ebXML (Electronic Business using XML) standards. For example, a direct trust relationship could be represented in an ebXML document; each federated domain that shares a direct trust relationship would have a copy of a contract that is expressed as an ebXML document. Operational characteristics for various entities within a federation may be specified within ebXML choreographies and published within ebXML registries; any enterprise that wishes to participate in a particular federation, e.g., to operate a trust proxy or trust broker, would need to conform to the published requirements that were specified by that particular federation for all trust proxies or trust brokers within the federation. A security token service could parse these ebXML documents for operational details on the manner in which tokens from other domains are to be translated. It should be noted, though, that other standards and mechanisms could be employed to support the present invention for specifying the details about the manner in which the trust relationships within a federation are implemented.

Single-Sign-on within Federated Architecture

During a given user's session, the user may visit many federated domains to use the web services that are offered by those domains. Domains can publish descriptions of services that they provide using standard specifications such as UDDI and WSDL, both of which use XML as a common data format. The user finds the available services and service providers through applications that also adhere to these standard

22

specifications. SOAP provides a paradigm for communicating requests and responses that are expressed in XML. Entities within a federated environment may employ these standards among others.

Within a federation, a user expects to have a single-sign-on experience in which the user completes a single authentication operation, and this authentication operation suffices for the duration of the user's session, regardless of the federation partners visited during that session. A session can be defined as the set of transactions from (and including) the initial user authentication, i.e. logon, to logout. Within a session, a user's actions will be governed in part by the privileges granted to the user for that session.

The federated architecture that is described hereinabove supports single-sign-on operations. To facilitate a single-sign-on experience, web services that support the federated environment will also support using an authentication assertion or security token generated by a third-party to provide proof of authentication of a user. This assertion will contain some sort of evidence of the user's successful authentication to the issuing party together with an identifier for that user. For example, a user may complete traditional authentication operation with one federation partner, e.g., by providing a username and password that the federation partners uses to build authentication credentials for the user, and then the federation partner is able to provide a SAML authentication assertion that is generated by the authenticating/issuing party to a different federation partner.

The federated environment also allows web services or other applications to request web services, and these web services would also be authenticated. Authentication in a web services environment is the act of verifying the claimed identity of the web services request so that the enterprise can restrict access to authorized clients. A user who requests or invokes a web service would almost always be authenticated, so the need for authentication within a federated environment that supports the present invention is not any different from current requirements of web services for user authentication.

Authentication of users that are accessing the computational resources of an enterprise without participating in a federated session are not impacted by the presence of a federated infrastructure. For example, an existing user who authenticates with a forms-based authentication mechanism over HTTP/S to access non-federated resources at a particular domain is not affected by the introduction of support at the domain for the federated environment. Authentication is handled in part by a point-of-contact server, which in turn may invoke a separate trust proxy or trust service component; the use of a point-of-contact server minimizes the impact on the infrastructure of an existing domain. For example, the point-of-contact server can be configured to pass through all non-federated requests to be handled by the back-end or legacy applications and systems at the domain.

The point-of-contact server may choose to invoke an HTTP-based authentication method, such as basic authentication, forms-based authentication, or some other authentication method. The point-of-contact server also supports a federation domain by recognizing an assertion that has been presented by the user as proof of authentication, such as an SAML authentication assertion, wherein the assertion has crossed between enterprise domains. The point-of-contact server may invoke the trust service, which in turn may invoke its security token service for validation of authentication credentials/security tokens.

Authentication of web services or other applications comprises the same process as authentication of users. Requests from web services carry a security token containing an

US 7,631,346 B2

23

authentication assertion, and this security token would be validated by the trust service in the same manner as a token presented by a user. A request from a web service should be accompanied by this token because the web service would have discovered what authentication assertions/security tokens were required by the requested service as advertised in UDDI.

With reference now to FIG. 6, a block diagram depicts a federated environment that supports federated single-sign-on operations. User 600, through a client device and an appropriate client application, such as a browser, desires to access a web service that is provided by enterprise/domain 610, which supports data processing systems that act as a federated domain within a federated environment. Domain 610 supports point-of-contact server 612 and trust proxy or trust service 614; similarly, domain 620 supports point-of-contact server 622 and trust proxy or trust service 624, while domain 630 supports point-of-contact server 632 and trust proxy or trust service 634. The trust proxies/services rely upon trust broker 650 for assistance, as described above. Additional domains and trust proxies/services may participate in the federated environment. FIG. 6 is used to describe a federated single-sign-on operation between domain 610 and domain 620; a similar operation may occur between domain 610 and domain 630.

The user completes an authentication operation with respect to domain 610; this authentication operation is handled by point-of-contact server 612. The authentication operation is triggered when the user requests access to some resource that requires an authenticated identity, e.g., for access control purposes or for personalization purposes. Point-of-contact server 612 may invoke a legacy authentication service, or it may invoke trust proxy 614 to validate the user's presented authentication credentials. Domain 610 becomes the user's identity provider or home domain for the duration of the user's federated session.

At some later point in time, the user initiates a transaction at a federation partner, such as enterprise 620 that also supports a federated domain, thereby triggering a federated single-sign-on operation. For example, a user may initiate a new transaction at domain 620, or the user's original transaction may cascade into one or more additional transactions at other domains. As another example, the user may invoke a federated single-sign-on operation to a resource in domain 620 via point-of-contact server 612, e.g., by selecting a special link on a web page that is hosted within domain 610 or by requesting a portal page that is hosted within domain 610 but that displays resources hosted in domain 620. Point-of-contact server 612 sends a request to trust proxy 614 to generate a federation single-sign-on token for the user that is formatted to be understood or trusted by domain 620. Trust proxy 614 returns this token to point-of-contact server 612, which sends this token to point-of-contact server 622 in domain. Domain 610 acts as an issuing party for the user at domain 620, which acts as a relying party. The user's token would be transferred with the user's request to domain 620; this token may be sent using HTTP redirection via the user's browser, or it may be sent by invoking the request directly of point-of-contact server 622 (over HTTP or SOAP-over-HTTP) on behalf of the user identified in the token supplied by trust proxy 614.

Point-of-contact server 622 receives the request together with the federation single-sign-on token and invokes trust proxy 624. Trust proxy 624 receives the federation single-sign-on token, validates the token, and assuming that the token is valid and trusted, generates a locally valid token for the user. Trust proxy 624 returns the locally valid token to point-of-contact server 622, which establishes a session for

24

the user within domain 620. If necessary, point-of-contact server 622 can initiate a federated single-sign-on at another federated partner.

Validation of the token at domain 620 is handled by the trust proxy 624, possibly with assistance from a security token service. Depending on the type of token presented by domain 610, the security token service may need to access a user registry at domain 620. For example, domain 620 may provide a binary security token containing the user's name and password to be validated against the user registry at domain 620. Hence, in this example, an enterprise simply validates the security token from a federated partner. The trust relationship between domains 610 and 620 ensures that domain 620 can understand and trust the security token presented by domain 610 on behalf of the user.

Federated single-sign-on requires not only the validation of the security token that is presented to a relying domain on behalf of the user but the determination of a locally valid user identifier at the relying domain based on information contained in the security token. One result of a direct trust relationship and the business agreements required to establish such a relationship is that at least one party, either the issuing domain or the relying domain or both, will know how to translate the information provided by the issuing domain into an identifier valid at the relying domain. In the brief example above, it was assumed that the issuing domain, i.e. domain 610, is able to provide the relying domain, i.e. domain 620, with a user identifier that is valid in domain 620. In that scenario, the relying domain did not need to invoke any identity mapping functionality. Trust proxy 624 at domain 620 will generate a security token for the user that will "vouch-for" this user. The types of tokens that are accepted, the signatures that are required on tokens, and other requirements are all pre-established as part of the federation's business agreements. The rules and algorithms that govern identifier translation are also pre-established as part of the federation's business agreements. In the case of a direct trust relationship between two participants, the identifier translation algorithms will have been established for those two parties and may not be relevant for any other parties in the federation.

However, it is not always the case that the issuing domain will know how to map the user from a local identifier for domain 610 to a local identifier for domain 620. In some cases, it may be the relying domain that knows how to do this mapping, while in yet other cases, neither party will know how to do this translation, in which case a third party trust broker may need to be invoked. In other words, in the case of a brokered trust relationship, the issuing and relying domains do not have a direct trust relationship with each other. They will, however, have a direct trust relationship with a trust broker, such as trust broker 650. Identifier mapping rules and algorithms will have been established as part of this relationship, and the trust broker will use this information to assist in the identifier translation that is required for a brokered trust relationship.

Domain 620 receives the token that is issued by domain 610 at point-of-contact server 622, which invokes trust proxy 624 to validate the token and perform identity mapping. In this case, since trust proxy 624 is not able to map the user from a local identifier for domain 610 to a local identifier for domain 620, trust proxy 624 invokes trust broker 650, which validates the token and performs the identifier mapping. After obtaining the local identifier for the user, trust proxy 624, possibly through its security token service, can generate any local tokens that are required by the back-end applications at domain 620, e.g., a Kerberos token may be required to facilitate single-sign-on from the point-of-contact server to the

US 7,631,346 B2

25

application server. After obtaining a locally valid token, if required, the point-of-contact server is able to build a local session for the user. The point-of-contact server will also handle coarse-grained authorization of user requests and forward the authorized requests to the appropriate application servers within domain **620**.

Federated User Lifecycle Management

A portion of the above description of FIGS. 2-6 explained an organization of components that may be used in a federated environment while other portions explained the processes for supporting single-sign-on operations across the federated environment. Service providers or relying domains within a federated environment do not necessarily have to manage a user's authentication credentials, and those relying domains can leverage a single single-sign-on token that is provided by the user's identity provider or home domain. The description of FIGS. 2-6 above, though, does not explain an explicit process by which federated user lifecycle management may be accomplished in an advantageous manner at the federated domains of federation partners.

Federated user lifecycle management functionality/service comprises functions for supporting or managing federated operations with respect to the particular user accounts or user profiles of a given user at multiple federated domains; in some cases, the functions or operations are limited to a given federated session for the user. In other words, federated user lifecycle management functionality refers to the functions that allow management of federated operations across a plurality of federated partners, possibly only during the lifecycle of a single user session within a federated computing environment.

Each federated domain might manage a user account, a user profile, or a user session of some kind with respect to the functions at each respective federated domain. For example, a particular federated domain might not manage a local user account or user profile within the particular federated domain, but the federated domain might manage a local user session for a federated transaction after the successful completion of a single-sign-on operation at the federated domain. As part of the federated user lifecycle management functionality that is supported by that particular federated domain, the federated domain can participate in a single-sign-off operation that allows the federated domain to terminate the local user session after the federated transaction is complete, thereby improving security and promoting efficient use of resources.

In another example of the use of federated user lifecycle management functionality, a user may engage in an online transaction that requires the participation of multiple federated domains. A federated domain might locally manage a user profile in order to tailor the user's experience with respect to the federated domain during each of the user's federated sessions that involve the federated domain. As part of the federated user lifecycle management functionality that is supported by that particular federated domain, the information in the federated domain's local user profile can be used in a seamless manner during a given federated transaction with information from other profiles at other federated domains that are participating in the given federated transaction. For example, the information from the user's multiple local user profiles might be combined in some type of merging operation such that the user's information is visually presented to the user, e.g., within a web page, in a manner such that the user is not aware of the different origins or sources of the user's information.

Federated user lifecycle management functionality may also comprise functions for account linking/delinking. A user is provided with a common unique user identifier across

26

federation partners, which enables single-sign-on and the retrieval of attributes (if necessary) about a user as part of the fulfillment of a request at one federation partner. Furthermore, the federation partner can request additional attributes from an identity provider using the common unique user identifier to refer to the user in an anonymous manner.

With reference now to FIG. 7, a block diagram depicts some of the components in a federated domain for implementing federated user lifecycle management functionality in order to support the present invention. FIG. 7 depicts elements at a single federated domain, such as the federated domain that is shown in FIG. 3. Some of the elements in FIG. 7 are similar or identical to elements that have been discussed hereinabove with respect to other figures, such as FIG. 3: point-of-contact server/service **702** is equivalent to point-of-contact server **342**; application servers **704**, which run services that control access to protected resources, are equivalent to application servers **334**; protected or controlled resources **706** are equivalent to protected resources **335**; and federated user lifecycle management (FULM) application **708** is equivalent to federated user lifecycle management application **352**. Although firewalls were not illustrated within FIG. 3, firewalls are illustrated within FIG. 7. Firewall **710** and firewall **712** create an external DMZ (electronic DeMilitarized Zone) that protects the enterprise's computing environment from computing threats outside of the enterprise's domain, e.g., via the Internet.

The different perspectives that are shown in FIG. 3 and FIG. 7 are not incompatible or at cross-purposes. In contrast with the example that is shown in FIG. 7, FIG. 3 does not illustrate the firewalls, yet point-of-contact server **342** resides within federation front-end **340**; in addition, federated user lifecycle management application **352** is contained within federation front-end **340**. In FIG. 7, point-of-contact server **702** is illustrated as residing within the DMZ between firewalls **710** and **712**, which form an electronic or physical front-end to the enterprise domain; in addition, federated user lifecycle management application/service **708** resides electronically behind firewall **712**. Trust service **714**, single-sign-on protocol service **716**, and identity-and-attribute service **718** employ enterprise user registry **720** and federation user registry **722** as necessary. The different perspectives of FIG. 3 and FIG. 7 can be reconciled by regarding federation front-end **340** and back-end **330** in FIG. 3 as a logical organization of components while regarding the DMZ and the other components in FIG. 7 as forming a physical or electronic front-end and a physical or electronic back-end, either of which may contain federated components.

Reiterating the roles of a point-of-contact entity/service, the point-of-contact entity provides session management, at least with respect to a user's interaction with the federation functionality with an enterprise's computing environment; applications within a legacy back-end of the enterprise's computing environment may also implement its own session management functionality. Assuming that an enterprise implements policy functionality with respect to the federated computing environment, the point-of-contact entity may act as a policy enforcement point to some other federation partner's policy decision point. In addition, assuming that it is permissible given the implementation of the federation functionality, the point-of-contact entity is responsible for initiating a direction authentication operation against a user in those scenarios in which a single-sign-on operation is not employed. As such, the point-of-contact entity may be implemented in a variety of forms, e.g., as a reverse proxy server, as a web server plug-in, or in some other manner. The point-of-contact functionality may also be implemented within an

US 7,631,346 B2

27

application server itself, in which case the federated user lifecycle management services may be logically located within the DMZ.

More importantly, referring again to FIG. 7, federated user lifecycle management application 708 also comprises support for interfacing to, interacting with, or otherwise interoperating with federated user lifecycle management plug-ins 724, which are not shown in FIG. 3. In the exemplary architecture that is shown in FIG. 7, federated protocol runtime plug-ins provide the functionality for various types of independently published or developed federated user lifecycle management standards or profiles, such as: WS-Federation Passive Client; and Liberty Alliance ID-FF Single Sign On (B/A, B/P and LECP), Register Name Identifier, Federation Termination Notification, and Single Logout. Different sets of federated protocols may be accessed at different URI's. This approach allows the federated user lifecycle management application to concurrently support multiple standards or specifications of federated user lifecycle management, e.g., the WS-Federation web services specification versus the Liberty Alliance's specifications, within a single application, thereby minimizing the configuration impact on the overall environment for supporting different federation protocols.

More specifically, the appropriate federated user lifecycle management functionality is invoked by the point-of-contact server by redirecting and/or forwarding user requests to the federated user lifecycle management application as appropriate. Referring again to FIG. 7, point-of-contact server 702 receives user requests 730, which are then analyzed to determine the type of request that has been received, which might be indicated by the type of request message that has been received or, as noted above, by determining the destination URI within the request message. While requests 732 for protected resources continue to be forwarded to application servers 704, requests 734 for federated user lifecycle management functions, e.g., requests to invoke a single-sign-off operation, are forwarded to federated user lifecycle management application 708, which invokes the appropriate federated user lifecycle management plug-in as necessary to fulfill the received request. When a new federation protocol or a new federated function is defined, or when an existing one is somehow modified or refined, support can be added simply by plugging a new support module or can be refined by modifying a previously installed plug-in.

The exemplary implementation of a federated user lifecycle management application in FIG. 7 illustrates that the federated user lifecycle management application is able to support multiple, simultaneous, federated user lifecycle management functions while providing a pluggability feature, thereby allowing new functionality to be added to the federated user lifecycle management application in the form of a plug-in when needed without requiring any changes to the existing infrastructure. For example, assuming that the present invention is implemented using a Java™-based federated user lifecycle management application, support for a new federation protocol, such as a newly published single-sign-on protocol, can be added by configuring newly developed Java™ classes to the Java™ CLASSPATH of the federated user lifecycle management application, wherein these new classes support the new standard along with the protocol interface for supporting the present invention.

The exemplary federated architecture leverages the existing environment in which a federated user lifecycle management solution is to be integrated. The federated user lifecycle management application can be easily modified to support new protocols/standards as they evolve with minimal changes to the overall infrastructure. Any changes that might be

28

required to support new federated user lifecycle management functionality are located almost exclusively within the federated user lifecycle management application, which would require configuring the federated user lifecycle management application to understand the added functionality.

There may be minimal configuration changes in other federated components, e.g., at a point-of-contact server, in order to allow the overall infrastructure to be able to invoke new federated user lifecycle management functionality while continuing to support existing federated user lifecycle management functionality. However, the federated user lifecycle management applications are functionally independent from the remainder of the federated components in that the federated user lifecycle management applications may require only minimal interaction with other federated components of the federated environment. For example, in an exemplary embodiment, the federated user lifecycle management functionality may integrate with an enterprise-based datastore, e.g., an LDAP datastore, if federated user lifecycle management information, such as NameIdentifier values in accordance with the Liberty Alliance profiles, are to be stored in an externally-accessible federated user lifecycle management datastore as opposed to a private, internal, federated user lifecycle management datastore that is not apparent or accessible to external entities.

Hence, an existing environment needs minimal modifications to support federated user lifecycle management functionality. Moreover, changes to federated user lifecycle management functionality, including the addition of new functionality, have minimal impact on an existing federated environment. Thus, when a new single-sign-on standard is published, support for this standard is easily added.

Traditional user authentication involves interaction between an enterprise's computing environment and the end-user only; the manner in which the enterprise chooses to implement this authentication interchange is the choice of the enterprise, which has no impact on any other enterprise. When federation or cross-domain single-sign-on functionality is desired to be supported, however, it becomes a requirement that enterprise partners interact with each other. This requirement cannot be done scalably using proprietary protocols. Although adding support for standards-based federation protocols directly to a point-of-contact entity seems like a robust solution, the point-of-contact entity, which is already an existing component within the enterprise's computing environment, must be modified; moreover, it must be modified every time that one of these public federation protocols changes. Moving this functionality out of the point-of-contact entity provides a more modular approach, wherein this pluggable functionality makes it easy to maintain migrations or updates to these protocols.

Runtime Linked-User-Account Creation During a Single-Sign-On Operation

With reference now to FIG. 8, a dataflow diagram depicts a typical prior art HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider. Although the processes that are illustrated within FIG. 8 and the subsequent figures employ HTTP-based communications, the present invention is not limited to HTTP-based communications, and other communication protocols may be employed; in particular, the present invention is not limited to front-channel communications as represented by HTTP redirection-based techniques but may be equally applied to back-channel techniques, such as SOAP/HTTP or SOAP/MQ. In the dataflow in FIG. 8, the user of a client, also known as a user agent, such as a web browser application, has

US 7,631,346 B2

29

already established a user account not only at the identity provider but also at the service provider (step 802) and that the user has already authenticated to the identity provider (step 804).

One of the prerequisites for the exemplary dataflow in FIG. 8 is that, at a minimum, that the user already has a federated account with the service provider; in other words, the service provider is required to recognize user identity information for the user when it receives this information from an identity provider in order to perform a single-sign-on operation for the user when initiated by an identity provider. With step 804, the prerequisite is simply that the user has authenticated to the identity provider at some previous point in time and currently has a valid session with the identity provider; there are no requirements on the reasons for which a session was established at step 804. It should be noted that other scenarios are possible, e.g., in which the identity provider determines at some later point in time after some interaction with the user that the identity provider only performs certain operations for authenticated users. It should also be noted that other scenarios for initiating a single-sign-on process are possible; for example, a user could initiate a single-sign-on operation by requesting a protected resource at the service provider, e.g., by using a bookmarked URL within a browser application.

The single-sign-on process commences at the identity provider by sending to the client from the identity provider an offer to provide access to federated resources (step 806); the offer may be in the form of hyperlinks to resources at federated web sites or, more generally, at federated domains, and the offer may be made in the form of an HTTP response message in response to a previous HTTP request message from the client (not shown) to view a particular web page on the identity provider's web site. The user then selects one of the offered federated resources at service providers that are known to the identity provider (step 808); the selection may be facilitated by an HTTP request message from the client to the identity provider.

The identity provider builds a message for requesting access to the selected federated resource on behalf of the user such that the message also includes a single-sign-on request (step 810), and the identity provider sends the resource request with the single-sign-on request to the client (step 812), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). The redirect message redirects the client to the appropriate location, e.g., as identified by a URI within the "Location" header of the redirect message, that identifies the appropriate service at the service provider that controls access to the requested federated resource.

It should be noted that, in the prior art, a preferred manner of initiating a push-type single-sign-on operation is to include the single sign-on assertion information in a request initiated by the identity provider and sent to the service provider in response to a request triggered directly at the service provider. For example, in a computing environment that implements a single-sign-on operation as specified in the Liberty Alliance ID-FF 1.2 specifications, an AuthnResponse is built. In some scenarios, it is required that the identity provider redirect the client to a trigger URL at the service provider, whereafter the trigger URL will cause the service provider to build a message, e.g., an AuthnRequest message within the Liberty Alliance specifications, that is redirected to the identity provider, which in turn causes the identity provider to build response, e.g., an AuthnResponse; this is the means by which a push-based SSO is implemented with Liberty ID-FF 1.1.

The outgoing request message may comprise two distinct requests: one requested operation for a single-sign-on opera-

30

tion on behalf of the user of the client, and another requested operation for accessing the protected resource that has been selected by the user of the client. Alternatively, the request to access the federated resource may include an implicit request to perform a single-sign-on operation for the user such that the single-sign-on process is merely part of a larger process for accessing the selected resource. The manner in which information is provided for the requested single-sign-on operation may vary. For example, the "Location" HTTP header of the redirect message may also include a query component, e.g., appended to a URI and demarcated within the URI with a "?" character, that contains various information, including security tokens for accomplishing the single-sign-on operation.

In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the appropriate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 814). In this manner, the client accesses the appropriate service at the service provider because the URI in the HTTP Get message still contains the attached information that is required by the service at the service provider.

The service provider receives the request message from the client, and assuming that the single-sign-on aspect of the request is successfully completed (step 816) such that the user has an active session at the service provider, the service provider then processes the resource access aspect of the request message (step 818). After processing the request message, the service provider responds by sending an HTTP response message to the client (step 820), thereby concluding the process.

With reference now to FIGS. 9A-9B, dataflow diagrams depict an HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention. In a manner similar to that shown in FIG. 8, both FIGS. 9A-9B depict a process by which an identity provider may request a single-sign-on operation at a selected service provider. However, whereas FIG. 8 depicts only a generalized single-sign-on operation at step 816, the processes that are shown in FIGS. 9A-9B differs from the process that is shown in FIG. 8 with respect to the single-sign-on operation. Unlike the process that is shown in FIG. 8, there is no prerequisite to the processes that are shown in FIGS. 9A-9B that the user of a client has already established a user account at the service provider.

More specifically, prior art solutions for single-sign-on operations have certain prerequisites, namely that the service provider and the identity provider must both recognize the user, and they must have an agreed-upon means of referring to this user, i.e. an alias identifier, or more simply, an alias; both prerequisites must be true before the initiation of a single-sign-on operation in order for the single-sign-on operation to be successful, otherwise it would fail. In these prior art solutions, the alias is included in the single-sign-on response that is sent from the identity provider to the service provider and used by the service provider to identify the user and build a session for the user. These prerequisites exist within prior art solutions whether: (a) the single-sign-on operation is triggered by the service provider, e.g., when a user accesses a protected resource that is hosted at the service provider and the service provider needs a session for the user, thereby initiating a single-sign-on operation by sending a request to an identity provider; or (b) the single-sign-on operation is triggered by the identity provider, e.g., when an identity provider has a list of linked resources that are hosted by related

US 7,631,346 B2

31

service providers, and after a user selects one of these links, a single-sign-on operation is initiated by the identity provider. In contrast, the present invention eliminates these prerequisites, as shown by multiple exemplary embodiments of the present invention that are described hereinbelow.

Referring to FIG. 9A, the user has already authenticated to the identity provider (step 902). With step 902, the user has authenticated to the identity provider at some previous point in time and currently has a valid session with the identity provider; there are no requirements on the reasons for which a session was established at step 902. It should be noted that other scenarios for initiating a single-sign-on operation may have a different sequence of steps. For example, a user might browse information that is provided by an identity provider without an authenticated session; at some later point in time after some interaction with the user, the user requests a resource such that the identity provider determines that the identity provider only performs certain operations for authenticated users, thereafter initiating an authentication operation with the user.

The single-sign-on process commences at the identity provider by sending to the client from the identity provider an offer to provide access to federated resources (step 904); the offer may be in the form of hyperlinks to resources at federated web sites or, more generally, at federated domains, and the offer may be made in the form of an HTTP response message in response to a previous HTTP request message from the client (not shown) to view a particular web page on the identity provider's web site. The user then selects one of the offered federated resources at service providers that are known to the identity provider (step 906); the selection may be accomplished by an HTTP request message from the client to the identity provider.

If the user does yet have a federated identity that may be used for a federated single-sign-on operation, then the identity provider creates an alias for the user (step 908). The identity provider builds a message for requesting access to the selected federated resource on behalf of the user such that the message also includes a single-sign-on request (step 910), i.e. a push-type single-sign-on request. A push-type single-sign-on request originates with an identity provider and is pushed to a service provider in an unsolicited manner in order to provide the service provider with information that authenticates a user identity; in contrast, a pull-type single-sign-on request originates with a service provider that is attempting to pull authentication information for a user in a solicited manner. Alternatively, a push-type single-sign-on operation can be emulated, particularly when a push-type single-sign-on operation is not supported explicitly. An emulated push-type single-sign-on operation occurs when an identity provider issues a request to a service provider, e.g., via a specialized service at the service provider such as an intersite transfer service, whereupon the service provider issues a single-sign-on request to the identity provider; after the identity provider responds to the request from the service provider, the processing steps are the same as if an explicit push-type single-sign-on operation has occurred.

The identity provider sends the resource request with the single-sign-on request to the client (step 912), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). The redirect message redirects the client to the appropriate location, e.g., as identified by a URI within the "Location" header of the redirect message, that identifies the appropriate service at the service provider that controls access to the requested federated resource. In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the appro-

32

priate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 914).

The service provider receives and processes the received request message from the client (step 916). At step 916, the service provider retrieves from the received message an alias identifier or alias information that has been associated with the user by the identity provider but which is not recognized by the service provider as being associated with a previously existing user account at the service provider. Hence, at step 916, the service provider determines that the user does not have a user account that links the user with the identity provider, i.e. a linked user account that informs the service provider that the service provider should accept information for a single-sign-on operation from the identity provider in order to authenticate the user. In other words, at step 916, the service provider determines that the service provider does not have a user account that links a user account at the service provider with a user account at the identity provider.

This recognition is significant for the following reasons. It should be noted that the user may already have one or more accounts at the service provider. These accounts may be used independently because a unique user account is based on a unique user identifier; given a user identifier, the service provider determines the privileges that are associated with the user identifier, i.e. a particular user account. Given that a linked user account is independent of any other user account that the user may have at the service provider, and given that a linked user account requires a unique user identifier to be associated with the linked user account, a linked user account is based on a user identifier that is independent and unique in comparison with any other user identifier that is known to the service provider; this particular user identifier is known as an alias identifier, although some form of alias information within multiple data variables may be used in place of a single alias data variable in some embodiments.

Therefore, after the service provider recognizes that a provided alias identifier is not associated with a previously existing user account at the service provider, the service provider begins to perform operations in order to ensure that the single-sign-on operation is performed successfully. Since the user had not yet been federated with the service provider, the service provider creates a new account for the user with the alias information that has been provided by the identity provider within the request message (step 918) such that the user has an active session at the service provider.

It should be noted that the minimum information required to create this account will be any local information that is required to identify the account (which is generated internally by the service provider) and any alias information that is provided by the identity provider; thereafter, the newly created user account is linked to the identity provider based on the provided alias for the user such that the service provider is able to perform a single-sign-on operation on behalf of the user in cooperation with the identity provider. If the request to the service provider includes user attributes as well as an alias identifier for the user, then these attributes may be added to the local account; however, it should be noted that, in some embodiments, the service provider may create a linked user account using only the alias identifier without any additional user attribute information from the identity provider, possibly by assigning a local set of default user attributes that are determined by the service provider and that are independent of any information that is received from the identity provider. Whether any additional attributes are provided or not, what should be noted is that this account would preferably not be enabled for direct authentication; hence, the only manner for

US 7,631,346 B2

33

the user to gain access to resources at the service provider would be as the result of a session established from a single-sign-on operation triggered by the identity provider.

After the linked user account has been created, the service provider then performs the requested resource access (step 920). After performing the resource access, the service provider responds by sending an HTTP response message to the client (step 922), thereby concluding the process.

The single-sign-on operation of the present invention, e.g., the embodiment that is shown in FIG. 9A, differs from the single-sign-on solutions of the prior art, e.g., the operation that is shown in FIG. 8, because the service provider recognizes during the single-sign-on operation of the present invention that the service provider does not have a user account for the user that links the user to an account at an identity provider in order to support single-sign-on operations, yet with the present invention the service provider is able to dynamically perform operations to allow the single-sign-on operation to proceed. More specifically, the service provider does not have, e.g., within its user registries or databases, enough information that allows the service provider to determine the user's local identity, and therefore the user's privileges, for the single-sign-on operation; without this information, the service provider cannot determine the appropriate set of privileges to be given to the user and, therefore, the type of session/access control rights to give the user, if the service provider were to allow the single-sign-on operation to proceed. In the prior art, the service provider cannot automatically create an active session for the user and allow access to protected resources; with the present invention, the service provider dynamically performs a runtime linked-user-account creation operation at the service provider by creating a linked user account based on the user identity, and possibly attribute information, that has been provided by the identity provider to the service provider, e.g., as provided in a request that has been redirected from the identity provider through the client. The service provider is willing and able to perform such operations based on its trust relationship with the identity provider with their federated computing environment. In this manner, the single-sign-on request can be fulfilled by the service provider, which results in the creation of an active session for the user, and the request for access to the protected resource can proceed.

In a manner similar to that shown in FIG. 9A, FIG. 9B depicts a process by which an identity provider may request a single-sign-on operation at a selected service provider; similar elements in the figures are identified by similar reference numerals. However, whereas FIG. 9A depicts only a generalized runtime user account creation operation at step 918, the process that is shown in FIG. 9B differs from the process that is shown in FIG. 9A with respect to the runtime linked-user-account creation operation, which stretches over steps 930-942 in FIG. 9B. Unlike the process that is shown in FIG. 9A, in the process that is shown in FIG. 9B, the service provider is not able to immediately create, or to complete immediately the creation of, a user account at the service provider based on the information that has been provided by the identity provider to the service provider.

Referring now to FIG. 9B, the single-sign-on processing in FIG. 9B (step 930) differs because the service provider recognizes during the single-sign-on operation that the service provider does not have a pre-existing user account for the identity that is claimed or asserted in the single-sign-on request from the identity provider, i.e. a pre-existing user account that links the service provider to the identity provider on behalf of the user, and further that the information contained in the single-sign-on request is not sufficient to create

34

a valid account at the service provider. Thus, the service provider recognizes that it requires additional information about the user from the identity provider. More specifically, the service provider does not have sufficient user attribute information to create an active account for the user; for example, there may be attributes that are required by the service provider but were not included in the received single-sign-on request. Because of these additional requirements, the service provider is not yet able to create, or to completely create, whatever user account registry/database entries that it requires based on the information it has received.

In the embodiment that is shown in FIG. 9B, the service provider responds by sending an HTTP redirect message (HTTP Response message with status/reason code "302") to the client (step 932); the message from the service provider contains a request for additional user attribute information. The redirect message redirects the client to the identity provider using a return URI that was previously provided by the identity provider to the service provider. In response to receiving the redirect message from the service provider, the client sends an HTTP Get message to the identity provider as indicated by the URI in the HTTP redirect message from the service provider (step 934).

The identity provider then processes the received message (step 936) to build a response that contains user attribute information that is stored by the identity provider about the user of the client. The manner in which specific user attributes are identified may vary. For example, the user attribute information that is provided by the identity provider may include user attribute information that was explicitly requested by the service provider. Additionally or alternatively, the user attribute information that is provided by the identity provider may include user attribute information that was implicitly requested by the service provider, i.e. user attribute information that has been determined by the identity provider to be sufficient for performing a user account creation operation at the service provider. In addition, the identity provider may perform an operation to reconcile the received request for additional user attribute information with the identity provider's previous single-sign-on request to ensure that a service provider is not attempting to obtain user attributes without sufficient reason to do so.

After the identity provider builds the message at step 936, the identity provider sends the message with the additional user attributes to the client (step 938), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). The redirect message redirects the client to the appropriate location, e.g., as identified by a URI within the "Location" header of the redirect message, that identifies the appropriate service at the service provider; the appropriate location may have been provided by the service provider in its request for the additional user attributes. In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the appropriate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 940).

Given the additional user attribute information in the received message, the service provider performs a runtime linked-user-account creation operation at the service provider (step 942) by creating a linked user account based on the received user attribute information; in this manner, the single-sign-on request (and any subsequent single-sign-on requests) can be fulfilled by the service provider, which also creates an active session for the user, and the request for access to the protected resource can proceed at steps 920 and 924. In some embodiments, the service provider may preliminarily create

US 7,631,346 B2

35

the user account yet postpone the completion of the creation of the user account; in these embodiments, the service provider completes the creation of the user account at step 942. It should be noted that the processes are described in the context of an HTTP redirection-based attribute retrieval, but these processes could be implemented with a back-channel approach, such as a SOAP/HTTP-based SAML attribute query from the service provider to the identity provider.

With reference now to FIGS. 9C-9E, dataflow diagrams depict an HTTP-redirection-based single-sign-on operation that is initiated by a federated identity provider to obtain access to a protected resource at a federated service provider with alternative methods for obtaining user attributes by the federated service provider in accordance with an embodiment of the present invention. The processes that are shown in FIGS. 9C-9E differ from the processes that are shown in FIGS. 9A-9B primarily in the manner in which the service provider obtains user attributes during a single-sign-on operation; the processes that are shown in FIGS. 9A-9B and in FIGS. 9C-9E are both initiated by an identity provider. In addition, in the dataflow that is shown in FIG. 9B, a service provider performs the runtime linked-user-account creation operation at step 942; in contrast, in the dataflow that is shown in FIGS. 9C-9E, the service provider may perform the runtime linked-user-account creation operation over multiple steps by partially creating a linked user account and then later completing the runtime linked-user-account creation operation, as described in more detail hereinbelow.

Referring now to FIG. 9C, the process commences with a user browsing public resources that are hosted by an identity provider (step 952), i.e. resources that are not protected resources that require an authentication operation with respect to the user prior to accessing the resources. At some subsequent point in time, the user's client sends to the identity provider a request for access to a protected resource that requires an authentication operation (step 954); for example, the user might attempt to browse information that the identity provider maintains about resources at service providers within a federated computing environment in which the identity provider participates with service providers, i.e. the identity provider's federated partners. In response to the user request, the identity provider performs an authentication operation with the user (step 956). In this example, the identity provider thereafter offers links to resources at federated service providers (step 958), and the user then selects or initiates an operation to access a resource at a service provider (step 960).

If the user does yet have an alias identifier that may be used for a federated single-sign-on operation, then the identity provider creates an alias for the user (step 962), which may include performing other operations to associate the alias identifier with the user, particular to associate the alias identifier with other information that is associated with the user, e.g., user attribute information. The identity provider builds a message for requesting access to the selected federated resource on behalf of the user such that the message also includes a single-sign-on request (step 964), i.e. a push-type single-sign-on request. The identity provider sends the resource request with the single-sign-on request to the client (step 966), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). The redirect message redirects the client to the appropriate location, e.g., as identified by a URI within the "Location" header of the redirect message, that identifies the appropriate service at the service provider that controls access to the requested federated resource. In response to receiving the redirect message from the identity provider, the client sends

36

an HTTP Get message to the appropriate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 968).

The service provider then processes the single-sign-on response (step 970), which may include, e.g., extracting the newly created alias identifier for the user and extracting any user attribute information about the user that has been preliminarily provided by the identity provider to the service provider. The service provider attempts to create a new user account (step 972) but either is not able to immediately create or is not able to create a fully appropriate user account at the service provider based on the federated user identity information that has been provided by the identity provider to the service provider in the single-sign-on response. In other words, depending on the implementation, the service provider either fails to create or determines that it cannot create a user account at this point in time, or the service provider creates a user account with limited-time or limited-access privileges. More specifically, the service provider does not have, e.g., within the combination of its user registries or databases and the information that is initially provided by the identity provider, the information required to allow the service provider to determine the user's appropriate set of privileges; without this information, the service provider cannot determine the type of session/access control rights to give the user. Hence, the service provider recognizes during the single-sign-on operation that the service provider does not yet have a linked user account for the user while also recognizing that the service provider requires additional information about the user from the identity provider (step 974).

The manner in which the service provider pulls additional user attribute information from the identity provider to complete the user account creation operation may vary; two examples of variations are shown in FIG. 9D and in FIG. 9E. FIG. 9D illustrates a front-channel attribute retrieval operation, whereas FIG. 9E illustrates a back-channel attribute retrieval operation. Hence, the data flow diagram that is shown in FIG. 9C continues into either FIG. 9D or FIG. 9E; in both cases, the dataflow diagrams in FIG. 9D and FIG. 9E conclude with similar steps that are denoted with similar reference numerals.

Referring now to FIG. 9D, the service provider responds to its determination of its lack of user attribute information by sending an HTTP redirect message (HTTP Response message with status/reason code "302") to the client (step 976); the message from the service provider contains a request for additional user attribute information. The redirect message redirects the client to the identity provider using a return URI that was previously provided by the identity provider to the service provider. In response to receiving the redirect message from the service provider, the client sends an HTTP Get message to the identity provider as indicated by the URI in the HTTP redirect message from the service provider (step 978). The identity provider then processes the received message (step 980) to build a response that contains user attribute information that is stored by the identity provider about the user of the client. The manner in which specific user attributes are identified may vary. For example, the user attribute information that is provided by the identity provider may include user attribute information that was explicitly requested by the service provider. Additionally or alternatively, the user attribute information that is provided by the identity provider may include user attribute information that was implicitly requested by the service provider, i.e. user attribute information that has been determined by the identity provider to be sufficient for performing a user account creation operation at the service provider. In addition, the identity provider may

US 7,631,346 B2

37

perform an operation to reconcile the received request for additional user attribute information with the service provider's previous single-sign-on request to ensure that a service provider is not attempting to obtain user attributes without sufficient reason to do so.

After the identity provider builds the message at step 980, the identity provider sends the message with the user attributes to the client (step 982), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the appropriate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 984).

Given the user attribute information in the received message, the service provider performs or completes a user account creation operation at the service provider (step 986) based on the received user attribute information; in this manner, the single-sign-on request can be fulfilled by the service provider, which also creates an active session for the user, and the request for access to the protected resource can proceed. In both FIG. 9D and FIG. 9E, the service provider provides access to the originally requested protected resource (step 988), and the service provider sends an HTTP Response message to the client that contains information that is derived from accessing the protected resource (step 990), thereby concluding the process.

Referring now to FIG. 9E, the service provider responds to its determination of its lack of user attribute information by sending a SOAP message directly from the service provider to the identity provider (step 992); the message from the service provider contains a request for required user attribute information. The identity provider then processes the received message (step 994) to build a response that contains user attribute information that is stored by the identity provider about the user of the client; again, the manner in which specific user attributes are identified may vary as described above. The identity provider sends a SOAP response message with the required user attributes to the client (step 996). Given the user attribute information in the received message, the service provider performs or completes a runtime user account creation operation at the service provider (step 998) based on the received user attribute information; in this manner, the single-sign-on request can be fulfilled by the service provider, which also creates an active session for the user, and the request for access to the protected resource can proceed at steps 988 and 990.

With reference now to FIG. 10, a flowchart depicts a more detailed process for performing a runtime linked-user-account creation operation at a service provider during a single-sign-on operation that has been initiated by an identity provider. The flowchart that is shown in FIG. 10 depicts a service-provider-centered perspective for some of the processing that occurs at a service provider within the dataflow diagram that is shown in FIG. 9B.

The process commences when the service provider receives a request from an identity provider to access a protected resource at the service provider on behalf of a user of a client based on a single-sign-on operation (step 1002). It should be noted that the protected resource may be an endpoint that corresponds to the service provider's functionality for fulfilling a single-sign-on request; in other words, it is possible that the request from the identity provider is redirected directly to the known functionality to accomplish the single-sign-on operation because the user has not requested a particular back-end resource but simply overall access to the service provider. The service provider extracts a user identi-

38

fier from the received request message (step 1004) and makes a determination as to whether or not the user identifier is recognized (step 1006).

If the user identifier is not recognized by the service provider, then the service provider cannot create an active session with the appropriate security considerations for protected access to its hosted computational resources. The service provider extracts any user attribute information that might be embedded in the received message (step 1008), and a determination is made as to whether or not the service provider has sufficient information about the user to create a user account for the user at that time (step 1010), e.g., as would be required to generate an entry into a user registry or whatever operations are typically performed by the service provider to create a local user account for the user at the service provider.

If the service provider does not have sufficient information about the user to create a user account for the user, the service provider may send a request to the identity provider to obtain additional user attribute information (step 1012); this may be performed in a synchronous or asynchronous manner. The service provider subsequently receives additional user attributes from the identity provider (step 1014).

It should be noted that an embodiment of the present invention may be implemented such that a user attribute retrieval operation by a service provider may be performed with respect to an attribute information provider. For example, a service provider may send a request for user attributes to an attribute information provider rather than an identity provider. Alternatively, a service provider may send a request for user attributes to an identity provider, which subsequently enlists assistance by sending a request to an attribute information provider, thereby acting as an intermediate trusted agent on behalf of the service provider. Additional information about the usage of an attribute information provider within the context of a federated computing environment may be found in Blakley et al., "Method and system for user-determined attribute storage in a federated environment", U.S. Patent Application Publication US 2004/0128378 A1, published Jul. 1, 2004, based on U.S. patent application Ser. No. 10/334,605, filed Dec. 31, 2002, which has a common assignee with the present patent application and is hereby incorporated by reference.

After receiving the additional user attributes at step 1014, or based on any user attribute information that may have been within the original request as extracted at step 1008, the service provider performs a linked user account creation operation (step 1016); as noted above, the operations that are performed to create a linked user account for the user at the service provider may vary. After the linked user account has been created at the service provider, the service provider may be able to proceed with the processing of the original request.

The service provider then determines whether there is sufficient information for activating the newly created account for the user (step 1018). If the service provider does not yet have sufficient information for activating the user's account, then a determination is made as to whether the service provider has already made too many attempts to obtain user attributes for activating the user account (step 1020); if not, then the process branches back to step 1012 to make an attempt to obtain the necessary user attributes. Otherwise, the service provider performs some type of error handling (step 1022), which may entail sending an error response back to the identity provider.

If the service provider has sufficient information for activating the user's account at step 1018, then the service provider creates an active session for the user (step 1024), either based on a successful authentication of a recognized user

US 7,631,346 B2

39

identity at step 1006 or based on the newly created user identity at step 1016. The service provider then generates a response for the original request to access the protected resource (step 1026), and the response is sent by the service provider to the identity provider (step 1028), thereby concluding the process.

The exemplary dataflow that is described with respect to FIG. 9B or the exemplary process that is described with respect to FIG. 10 can be described in the context of the exemplary data processing systems that are shown in FIG. 7 or in FIG. 3.

In order to perform a single-sign-on operation, the identity and attribute service (I&AS) 356 or 718 needs to be able to recognize a user identity from a single-sign-on request within a user registry in some manner. The user registry can be a local registry, e.g., private to the installation of the federated functionality, such as federation user registry 358 or 720, or it could be an enterprise registry that is shared by other applications within an enterprise, such as enterprise user registry 338 or 722. The user account information or user registry information needs to allow I&AS 356 or 718 to build the appropriate information that identifies the user locally. This information may be represented by a username, a set of attributes, e.g., groups and/or roles and/or entitlements, or an opaque identifier, e.g., a NamelIdentifier as described within the Liberty Alliance specifications. If the information asserted by the identity provider about the user for which a single-sign-on is being requested cannot be found within a local registry, then I&AS 356 or 718 is not able to build information about the user in a local manner, e.g., to create valid credentials that are used by entities within the local enterprise; in addition, the point-of-contact server 342 or 702 is not able establish a session for the user, and the user is not able to access protected resources.

The present invention provides a mechanism to prevent the generation of some form of error code for an unrecognized user as would be performed in prior art approaches. In the embodiment of the present invention that is shown in the figures, I&AS 356 or 718 may attempt to create a user account, record, or entry, as appropriate for the data storage requirements of the enterprise when the received user identity information is unrecognized. I&AS 356 or 718 performs these operations based on the trust relationship between the federated partners of the identity provider and the service provider, e.g., as provided for within a configured policy that allows for this type of action. The user identity information that is received in the original request may be an actual username value or an opaque data value such as a Liberty Alliance NamelIdentifier; this user identity information may be used as a pointer into a user registry from which this user's information would be accessed at some subsequent point in time, or it may be used in a temporary manner until permanent data records are created, e.g., as a pointer into a cache that contains temporary information.

If the original single-sign-on request contains attribute data about a user, then the attribute information may be added directly to a user registry when creating a user account for the user locally. However, the originally received request does not necessarily contain all of the information that is required locally to create a user account for the user. In this case, I&AS 356 or 718 working with single-sign-on protocol service (SPS) 354 or 716 may issue an attribute query, such as a SAML attribute query and/or a WS-AttributeService request, to the user's identity provider to retrieve additional information about the user. As a result, the service provider is able to create a user account for the user in a runtime manner, i.e.

40

while the single-sign-on operation is suspended or, from another temporal viewpoint, concurrently with or as part of the single-sign-on operation.

The level of trust that is accorded to this type of runtime linked-user-account creation operation by the service provider may vary; it is not necessarily the case that all service providers would be willing to allow a completely automated, dynamically determined, linked-user-account creation operation. Hence, in some cases, a service provider may require that a local workflow operation must be undertaken, e.g., a type of local administrative approval process. Rather than directly creating a user account for a user in a completely automated process, the federated user lifecycle management (FULM) application/service 352 or 708 might store the user information, including additional, out-of-band-retrieved user attributes, into some form of local datastore, which may, in turn, trigger a local workflow/approval process. This requirement allows a administrative user at the service provider to approve the creation of the user account in accordance with local policy requirements, etc. As a result, the runtime user account creation operation may be asynchronous, in which case the service provider may send a message to the user to indicate that an account is being created for the user at the service provider and to indicate that the user can attempt to perform a login operation at the service provider at some subsequent point in time. If the user needs to be added to more than an I&AS-accessible user registry, then FULM service 352 or 708 may send this information to a local datastore, from which a local user account creation process is initiated; in this case, accounts/records can be created for the user in multiple destinations in addition to a local I&AS-accessible datastore. It should be noted that the service provider may choose to grant the user a limited-time, limited-access session, thereby restricting the user to a subset of accessible resources rather than a complete set of resources until some later event, e.g., completion of some type of workflow process for approving broader access, or until some later determination is made as to the privileges that should be given to the user; thereafter, the user account would be created with an appropriate level of access to resources.

With reference now to FIG. 11A, a dataflow diagram depicts an HTTP-redirection-based pull-type single-sign-on operation that is initiated by a federated service provider to allow access to a protected resource at the federated service provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention. The processes that are shown in FIGS. 11A-11D differ from the processes that are shown in FIGS. 9A-9B primarily in the perspective of the origination of the single-sign-on operation; the processes that are shown in FIGS. 9A-9B are initiated by an identity provider, whereas the processes that are shown in FIGS. 11A-11D are initiated by a service provider.

Referring now to FIG. 11A, the process commences with a user browsing public resources that are hosted by a service provider (step 1102), i.e. resources that are not protected resources that require an authentication operation with respect to the user prior to accessing the resources. At some subsequent point in time, the user's client sends to the service provider a request for access to a protected resource that requires an authentication operation (step 1104). The service provider recognizes that it does not have an identifier for the user's preferred identity provider, and the service provider prompts the user to provide the information in some manner (step 1106), e.g., through user-selectable controls in a web page. The client then sends to the service provider the appro-

US 7,631,346 B2

41

priate information about the preferred identity provider as selected by the user (step 1108).

The service provider then builds a pull-type single-sign-on request for the user (step 1110). In this example, the service provider assumes that the user is a federated user, e.g., by assuming that the identity provider maintains some type of federated alias identifier for the user. The service provider sends the single-sign-on request to the client (step 1112), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code “302”). In response to receiving the redirect message from the service provider, the client sends an HTTP Get message to the identity provider as indicated by the URI in the HTTP redirect message from the service provider (step 1114).

In response to receiving the pull-type single-sign-on request, the identity provider performs an authentication operation with the client with respect to the user of the client, if required (step 1116); for example, the authentication operation may not be required if the user is already logged on at the identity provider, i.e. if the user already has an active session at the identity provider. The identity provider evaluates the received request (step 1118), and the identity provider determines that the service provider has not provided a user identity to the identity provider as part of the federation request. If the identity provider further determines that the user does not yet have a federated identity, then the identity provider creates a federated identity for the user, e.g., an alias identifier for the user (step 1120). The identity provider builds a pull-type single-sign-on response (step 1122) that contains the newly created, federated identity for the user and optionally contains additional user attribute information that is managed by the identity provider about the user; this pull-type single-sign-on response may or may not have the same data format characteristics as a push-type single-sign-on response. The identity provider sends the single-sign-on response to the client (step 1124), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code “302”). In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 1126).

The service provider then processes the single-sign-on response (step 1128), which may include, e.g., extracting the newly created federated identifier for the user and may also include the extraction of additional user attribute information about the user. The service provider then creates a new linked user account for the user with the federated identifier that was provided by the identity provider (step 1130), which may possibly also entail using some user attribute information about the user from the identity provider if the identity provider has sent such information and if the service provider needs this information while creating the new user account. After the user has an active session based on the newly created user account, the service provider provides access to the originally requested protected resource (step 1132), and the service provider sends an HTTP Response message to the client that contains information that is derived from accessing the protected resource (step 1134), thereby concluding the process.

With reference now to FIGS. 11B-11D, a set of dataflow diagrams depict an HTTP-redirection-based pull-type single-sign-on operation that is initiated by a federated service provider to allow access to a protected resource at the federated service provider with additional retrieval of user attribute information from a federated identity provider while performing a runtime linked-user-account creation operation at the federated service provider in accordance with an embodiment of the present invention. In a manner similar to that shown in FIG. 11A, FIG. 11B depicts a process by which a service provider may request a single-sign-on operation at a selected identity provider; similar elements in the figures are identified by similar reference numerals. However, whereas FIG. 11A depicts only a generalized runtime linked-user-account creation operation at step 1130, the process that is shown in FIG. 11B differs from the process that is shown in FIG. 11A with respect to the runtime linked-user-account creation operation, which stretches over steps 1150-1164 in FIGS. 11A-11B.

42

Referring now to FIG. 11B, unlike the process that is shown in FIG. 11A, the service provider attempts to create a new user account (step 1150) but either is not able to immediately create or is not able to create a fully appropriate user account at the service provider based on the federated user identity information that has been provided by the identity provider to the service provider in the single-sign-on. In other words, depending on the implementation, the service provider either fails to create or determines that it cannot create a user account at this point in time, or the service provider creates a user account with limited-time or limited-access privileges. Hence, the single-sign-on processing in FIG. 11B differs because the service provider recognizes during the single-sign-on operation that the service provider does not yet have a linked user account for the user while also recognizing that the service provider requires additional information about the user from the identity provider (step 1152). More specifically, the service provider does not have, e.g., within its user registries or databases, enough information that allows the service provider to determine the user’s appropriate set of privileges; without this information, the service provider cannot determine the type of session/access control rights to give the user.

35

The manner in which the service provider pulls additional user attribute information from the identity provider to complete the linked-user-account creation operation may vary; two examples of variations are shown over steps 1154-1164 in FIG. 11C or over steps 1172-1178 in FIG. 11D. FIG. 11C illustrates a front-channel attribute retrieval operation, whereas FIG. 11D illustrates a back-channel attribute retrieval operation. Hence, the data flow diagram that is shown in FIG. 11B continues into either FIG. 11C or FIG. 11D; in both cases, the dataflow diagrams in FIG. 11C and FIG. 11D conclude with similar steps. In both FIG. 11C and FIG. 11D, the service provider provides access to the originally requested protected resource at step 1132, and the service provider sends an HTTP Response message to the client that contains information that is derived from accessing the protected resource at step 1134, thereby concluding the process.

40

Referring now to FIG. 1C, the service provider responds to its determination of its lack of user attribute information by sending an HTTP redirect message (HTTP Response message with status/reason code “302”) to the client (step 1154); the message from the service provider contains a request for additional user attribute information. The redirect message directs the client to the identity provider using a return URI that was previously provided by the identity provider to the service provider. In response to receiving the redirect message from the service provider, the client sends an HTTP Get message to the identity provider as indicated by the URI in the HTTP redirect message from the service provider (step 1156). The identity provider then processes the received message (step 1158) to build a response that contains user attribute information that is stored by the identity provider about the user of the client. The manner in which specific user

45

50

55

60

65

US 7,631,346 B2

43

attributes are identified may vary. For example, the user attribute information that is provided by the identity provider may include user attribute information that was explicitly requested by the service provider. Additionally or alternatively, the user attribute information that is provided by the identity provider may include user attribute information that was implicitly requested by the service provider, i.e. user attribute information that has been determined by the identity provider to be sufficient for performing a linked-user-account creation operation at the service provider. In addition, the identity provider may perform an operation to reconcile the received request for additional user attribute information with the service provider's previous single-sign-on request to ensure that a service provider is not attempting to obtain user attributes without sufficient reason to do so.

After the identity provider builds the message at step 1158, the identity provider sends the message with the additional user attributes to the client (step 1160), e.g., in the form of an HTTP redirect message (HTTP Response message with status/reason code "302"). In response to receiving the redirect message from the identity provider, the client sends an HTTP Get message to the appropriate service at the service provider as indicated by the URI in the HTTP redirect message from the identity provider (step 1162).

Given the additional user attribute information in the received message, the service provider performs or completes a runtime linked-user-account creation operation at the service provider (step 1164) based on the received user attribute information; in this manner, the single-sign-on request can be fulfilled by the service provider, which also creates an active session for the user, and the request for access to the protected resource can proceed at steps 1132 and 1134.

Referring now to FIG. 11D, the service provider responds to its determination of its lack of user attribute information by sending a SOAP message directly from the service provider to the identity provider (step 1172); the message from the service provider contains a request for additional user attribute information. The identity provider then processes the received message (step 1174) to build a response that contains user attribute information that is stored by the identity provider about the user of the client; again, the manner in which specific user attributes are identified may vary as described above. The identity provider sends a SOAP response message with the additional user attributes to the client (step 1176). Given the additional user attribute information in the received message, the service provider performs or completes a runtime linked-user-account creation operation at the service provider (step 1178) based on the received user attribute information; in this manner, the single-sign-on request can be fulfilled by the service provider, which also creates an active session for the user, and the request for access to the protected resource can proceed at steps 1132 and 1134.

CONCLUSION

The advantages of the present invention should be apparent in view of the detailed description of the invention that is provided above. When an identity provider attempts to initiate a single-sign-on operation on behalf of a user at a service provider in order to obtain access to a controlled resource that is hosted by the service provider, it is possible that the service provider would not recognize the user identifier or other user identity information in the received request. In the prior art, this scenario would generate an error.

With the present invention, the service provider may dynamically perform a linked-user-account creation opera-

44

tion as part of the single-sign-on operation. If necessary, the service provider can pull additional user attribute information from the identity provider in order to obtain the required user attributes for the user in a manner that is locally required by the identity management functionality of the service provider.

"It is important to note that while the present invention has been described in the context of a fully functioning data processing system, those of ordinary skill in the art will appreciate that the processes associated with the present invention are capable of being distributed in the form of instructions in a computer readable medium. Examples of computer readable media include media such as EPROM, ROM, tape, paper, floppy disc, hard disk drive, RAM, and CD-ROMS."

A method is generally conceived to be a self-consistent sequence of steps leading to a desired result. These steps require physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It is convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, parameters, items, elements, objects, symbols, characters, terms, numbers, or the like. It should be noted, however, that all of these terms and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities.

The description of the present invention has been presented for purposes of illustration but is not intended to be exhaustive or limited to the disclosed embodiments. Many modifications and variations will be apparent to those of ordinary skill in the art. The embodiments were chosen to explain the principles of the invention and its practical applications and to enable others of ordinary skill in the art to understand the invention in order to implement various embodiments with various modifications as might be suited to other contemplated uses.

What is claimed is:

1. A method for managing user authentication within a distributed data processing system, wherein a first system and a second system interact within a federated computing environment and support single-sign-on operations in order to provide access to protected resources, at least one of the first system and the second system comprising a processor, the method comprising:

triggering a single-sign-on operation on behalf of the user in order to obtain access to a protected resource that is hosted by the second system, wherein the second system requires a user account for the user to complete the single-sign-on operation prior to providing access to the protected resource;

receiving from the first system at the second system an identifier associated with the user; and

creating a user account for the user at the second system based at least in part on the received identifier associated with the user after triggering the single-sign-on operation but before generating at the second system a response for accessing the protected resource, wherein the created user account supports single-sign-on operations between the first system and the second system on behalf of the user.

2. The method of claim 1 further comprising:

creating an alias identifier for the user at the first system after triggering the single-sign-on operation.

3. The method of claim 1 further comprising:

sending a message from the second system to the first system to pull authentication information for the user

US 7,631,346 B2

45

from the first system to the second system in order to trigger the single-sign-on operation for the user at the second system.

4. The method of claim 1 further comprising:

receiving a message from the first system at the second system to push authentication information for the user from the first system to the second system in order to trigger the single-sign-on operation for the user at the second system.

5. The method of claim 1 further comprising:

in response to a determination at the second system that the second system does not have sufficient user attribute information to complete creation of a user account for the user at the second system, sending a request message from the second system to the first system to retrieve user attribute information; and

receiving at the second system from the first system a response message that contains user attribute information that is employed by the second system to complete creation of a user account for the user at the second system.

6. The method of claim 5 further comprising:

employing a front-channel information retrieval mechanism to obtain user attribute information.

7. The method of claim 6 further comprising:

using HyperText Transport Protocol (HTTP) in the front-channel information retrieval mechanism.

8. The method of claim 5 further comprising:

employing a back-channel information retrieval mechanism to obtain user attribute information.

9. The method of claim 8 further comprising:

using Simple Object Access Protocol (SOAP) in the back-channel information retrieval mechanism.

10. The method of claim 5 further comprising:

performing a preliminary user account creation operation to commence creation of the user account for the user prior to retrieving user attribute information for the user; and

performing a concluding user account creation operation to complete creation of the user account for the user after retrieving user attribute information for the user.

11. The method of claim 5 further comprising:

retrieving user attribute information from a fourth system by the first system.

12. The method of claim 1 wherein the first system supports an identity provider and the second system supports a service provider.

13. The method of claim 12 further comprising:

prompting the user by the service provider to provide or to select an identifier for the identity provider prior to receiving an identifier associated with the user.

14. The method of claim 1 further comprising:

in response to a determination at the second system that the second system does not have sufficient user attribute information to complete creation of a user account for the user at the second system, sending a request message to a fourth system to retrieve user attribute information; and

receiving at the second system from the fourth system a response message that contains user attribute information that is employed by the second system to complete creation of a user account for the user at the second system.

15. A computer program product on a computer-readable medium for managing user authentication within a data processing system, wherein a first system and a second system interact within a federated computing environment and sup-

10

25

30

45

60

65

46

port single-sign-on operations in order to provide access to protected resources, at least one of the first system and the second system comprising a processor, the computer program product holding computer program instructions which when executed by the data processing system perform a method comprising:

triggering a single-sign-on operation on behalf of the user in order to obtain access to a protected resource that is hosted by the second system, wherein the second system requires a user account for the user to complete the single-sign-on operation prior to providing access to the protected resource;

receiving from the first system at the second system an identifier associated with the user; and

creating a user account for the user at the second system based at least in part on the received identifier associated with the user after triggering the single-sign-on operation but before generating at the second system a response for accessing the protected resource, wherein the created user account supports single-sign-on operations between the first system and the second system on behalf of the user.

16. The computer program product of claim 15 wherein the method further comprises:

creating an alias identifier for the user at the first system after triggering the single-sign-on operation.

17. The computer program product of claim 15 wherein the method further comprises:

sending a request message from the second system to the first system to retrieve user attribute information in response to a determination at the second system that the second system does not have sufficient user attribute information to complete creation of a user account for the user at the second system; and

receiving at the second system from the first system a response message that contains user attribute information that is employed by the second system to complete creation of a user account for the user at the second system.

18. An apparatus for managing user authentication within a data processing system, wherein a first system and a second system interact within a federated computing environment and support single-sign-on operations in order to provide access to protected resources, at least one of the first system and the second system comprising a processor, the apparatus comprising:

a processor;

a computer memory holding computer program instructions which when executed by the processor perform a method comprising:

triggering a single-sign-on operation on behalf of the user in order to obtain access to a protected resource that is hosted by the second system, wherein the second system requires a user account for the user to complete the single-sign-on operation prior to providing access to the protected resource;

receiving from the first system at the second system an identifier associated with the user; and

creating a user account for the user at the second system based at least in part on the received identifier associated with the user after triggering the single-sign-on operation but before generating at the second system a response for accessing the protected resource, wherein the created user account supports single-sign-on operations between the first system and the second system on behalf of the user.

US 7,631,346 B2

47

19. The apparatus of claim **18** wherein the method further comprises:

creating an alias identifier for the user at the first system after triggering the single-sign-on operation.

20. The apparatus of claim **18** wherein the method further comprises:

sending a request message from the second system to the first system to retrieve user attribute information in response to a determination at the second system that the

5

48

second system does not have sufficient user attribute information to complete creation of a user account for the user at the second system; and

receiving at the second system from the first system a response message that contains user attribute information that is employed by the second system to complete creation of a user account for the user at the second system.

* * * * *

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 7,631,346 B2
APPLICATION NO. : 11/097587
DATED : December 8, 2009
INVENTOR(S) : Hinton et al.

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

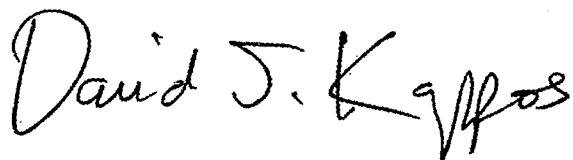
On the Title Page:

The first or sole Notice should read --

Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b)
by 1217 days.

Signed and Sealed this

Twenty-first Day of December, 2010



David J. Kappos
Director of the United States Patent and Trademark Office

EXHIBIT C



US008315904B2

(12) **United States Patent**
Black et al.

(10) **Patent No.:** US 8,315,904 B2
(45) **Date of Patent:** Nov. 20, 2012

(54) **ORGANIZATION FOR PROMOTION MANAGEMENT**

(75) Inventors: **Andre B. Black**, Medford, MA (US);
David F. Cheung, Potomac, MD (US);
Ruby Kennedy, Bedford, MA (US);
Yuchun Lee, Sudbury, MA (US);
Patrick Martin, Lexington, MA (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1069 days.

(21) Appl. No.: **11/492,010**

(22) Filed: **Jul. 24, 2006**

(65) **Prior Publication Data**

US 2008/0033808 A1 Feb. 7, 2008

6,067,525 A	5/2000	Johnson et al.
6,076,068 A	6/2000	DeLapa et al.
6,078,892 A	6/2000	Anderson et al.
6,097,792 A	8/2000	Thornton
6,134,530 A	10/2000	Bunting et al.
6,160,549 A	12/2000	Touma et al.
6,161,103 A	12/2000	Rauer et al.
6,263,334 B1	7/2001	Fayyad et al.
6,321,241 B1	11/2001	Gartung et al.
6,493,708 B1	12/2002	Ziauddin et al.
6,539,232 B2	3/2003	Hendrey et al.
6,567,786 B1	5/2003	Bibelnieks et al.
6,606,621 B2	8/2003	Hopeman et al.
6,684,207 B1	1/2004	Greenfield et al.
6,691,120 B1	2/2004	Durrant et al.
6,694,316 B1	2/2004	Langseth et al.
6,782,390 B2	8/2004	Lee et al.
6,847,934 B1	1/2005	Lin et al.
6,868,389 B1	3/2005	Wilkins et al.
6,915,289 B1	7/2005	Malloy et al.
6,917,940 B1	7/2005	Chen et al.
6,959,305 B2	10/2005	Bird et al.
2001/0014868 A1	8/2001	Herz et al.
2002/0026356 A1	2/2002	Bergh et al.
2002/0046116 A1	4/2002	Hohle et al.
2002/0169654 A1*	11/2002	Santos et al. 705/10

(Continued)

(51) **Int. Cl.**

G06Q 30/00 (2012.01)

(52) **U.S. Cl.** 705/14.42; 705/14.41

(58) **Field of Classification Search** 705/14.41,
705/14.42

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,664,171 A	9/1997	Agrawal et al.
5,765,154 A	6/1998	Horikiri et al.
5,822,751 A	10/1998	Gray et al.
5,883,940 A	3/1999	Thornton
5,899,988 A	5/1999	Depledge et al.
5,930,764 A	7/1999	Melchionne et al.
5,966,695 A	10/1999	Melchionne et al.
5,974,407 A	10/1999	Sacks
6,061,658 A	5/2000	Chou et al.

OTHER PUBLICATIONS

U.S. Appl. No. 09/575,283, filed May 22, 2000, Bergh et al.

(Continued)

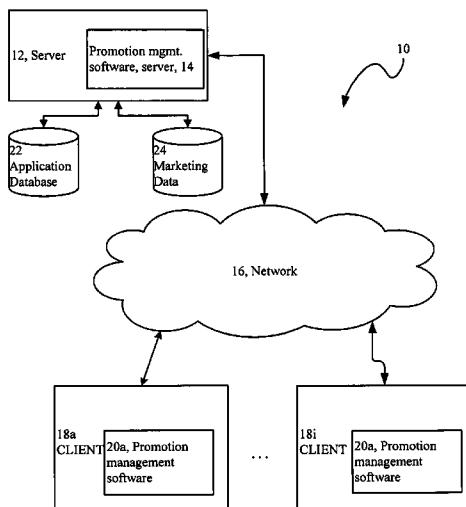
Primary Examiner — Daniel Sorkowitz

(74) **Attorney, Agent, or Firm** — Cuenot, Forsythe & Kim, LLC

(57) **ABSTRACT**

A computer implemented method for producing a promotion list for a promotion management campaign is described. The method includes assigning one or more promotion instances to the promotions list, and storing the promotion list in an electronic medium.

21 Claims, 26 Drawing Sheets



US 8,315,904 B2

Page 2

U.S. PATENT DOCUMENTS

2003/0009467 A1	1/2003	Perrizo
2003/0083924 A1	5/2003	Lee et al.
2003/0110119 A1	6/2003	Crites
2003/0115207 A1	6/2003	Bowman et al.
2003/0130883 A1*	7/2003	Schroeder et al.
2003/0135442 A1	7/2003	Kumar et al.
2003/0144873 A1	7/2003	Keshel
2003/0187767 A1	10/2003	Crites et al.
2003/0208402 A1*	11/2003	Bibelnieks et al.
2004/0133487 A1	7/2004	Hanagan et al.
2004/0254860 A1	12/2004	Wagner et al.
2005/0197899 A1*	9/2005	Veit et al.
2005/0216525 A1*	9/2005	Wachholz-Prill et al.
2005/0289000 A1	12/2005	Chiang et al.
2006/0085246 A1*	4/2006	Li
2006/0095321 A1*	5/2006	Shi et al.
2007/0150334 A1	6/2007	Bergh et al.
2008/0021909 A1	1/2008	Black et al.
2008/0033784 A1	2/2008	Chalimadugu et al.
2008/0033807 A1	2/2008	Black et al.
2008/0033809 A1	2/2008	Black et al.
2008/0126146 A1	5/2008	Benveniste et al.

OTHER PUBLICATIONS

- Griggs. "Give us leads! Give us leads!" Sales and Marketing Management. New York 149(7):66, 1997 (from Proquest).
- McGinnis et al., "Strategic account management in the new procurement environment", Supply Chain Management. Bradford 3(1):12, 1998 (from Proquest).

"General Interactive Teams up with IBM to Deliver Industries Best E-Business and Relationship Marketing Solutions". Business/Tech-nology Editors, Business Wire, New York. May 3, 1999 (from Proquest).

"Microsoft Announces Internet Lead Management Platform", from www.microsoft.com, Jan. 24, 2000.

"Data Mining Solutions and Customer Management", Bill Bradway, Meridien Research, Inc., Dec. 1997.

Godwin. "Predictor of Households' Debt Repayment Difficulties", Association for Financial Counseling and Planning Education, pp. 67-78, © 1999.

Boyes et al., "Lender Reactions to Information Restrictions: The Case of Banks and the ECOA", Journal of Money, Credit and Banking 18(2):211-219, 1986.

Chou et al., "Identifying Prospective Customers", © ACM, pp. 447-456, 2000.

Hernandez-Sherrington et al., "A Generalization of Band Joins and the Merge/Purge Problem", © Columbia University, Ph.D. Thesis, pp. i-171, 1996.

Raab. "Unica Impact!" DM News, Sep. 1999, retrieved from the Internet at URL: <http://www.raabassociates.com/a909unic.htm>.

"Leang Customer-Focused Corporations Join Unica TM Advisory Board", BusinessWire, pp. 1-2, Nov. 6, 2000.

"Lands' End to Automate Marketing Processes with Unica Corporation's Affinium", BusinessWire pp. 1-2, Sep. 12, 2000.

"Product Review 'Affinium Campaign'—Affinium Campaign Enables MITI to Create Complex Campaigns Across Diverse Data-bases", DM Review, pp. 1, May 2000.

"Workflow and Internet: Catalysts for Radical Change", Workflow Management Coalition, Jun. 1998, pp. 1-32.

* cited by examiner

U.S. Patent

Nov. 20, 2012

Sheet 1 of 26

US 8,315,904 B2

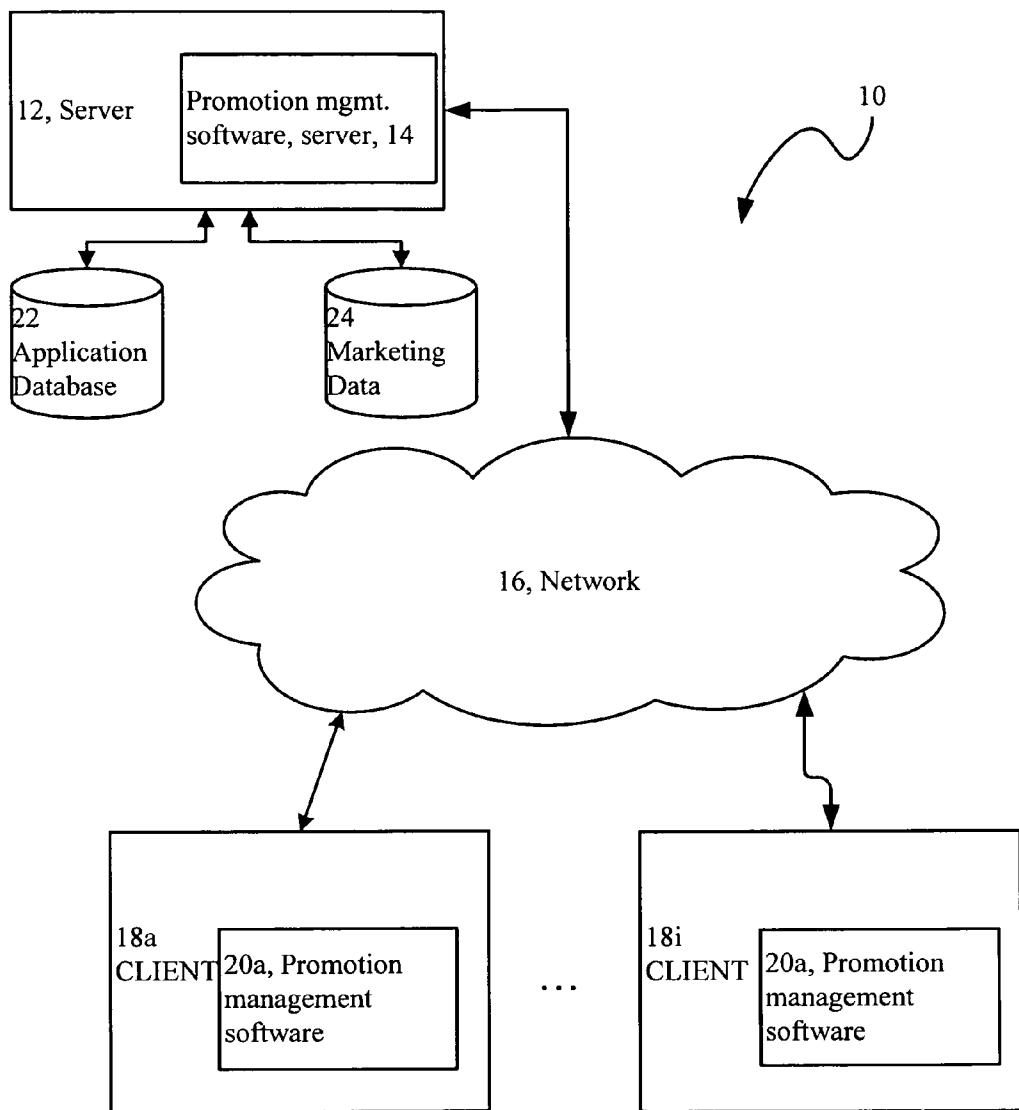


FIG. 1

U.S. Patent

Nov. 20, 2012

Sheet 2 of 26

US 8,315,904 B2

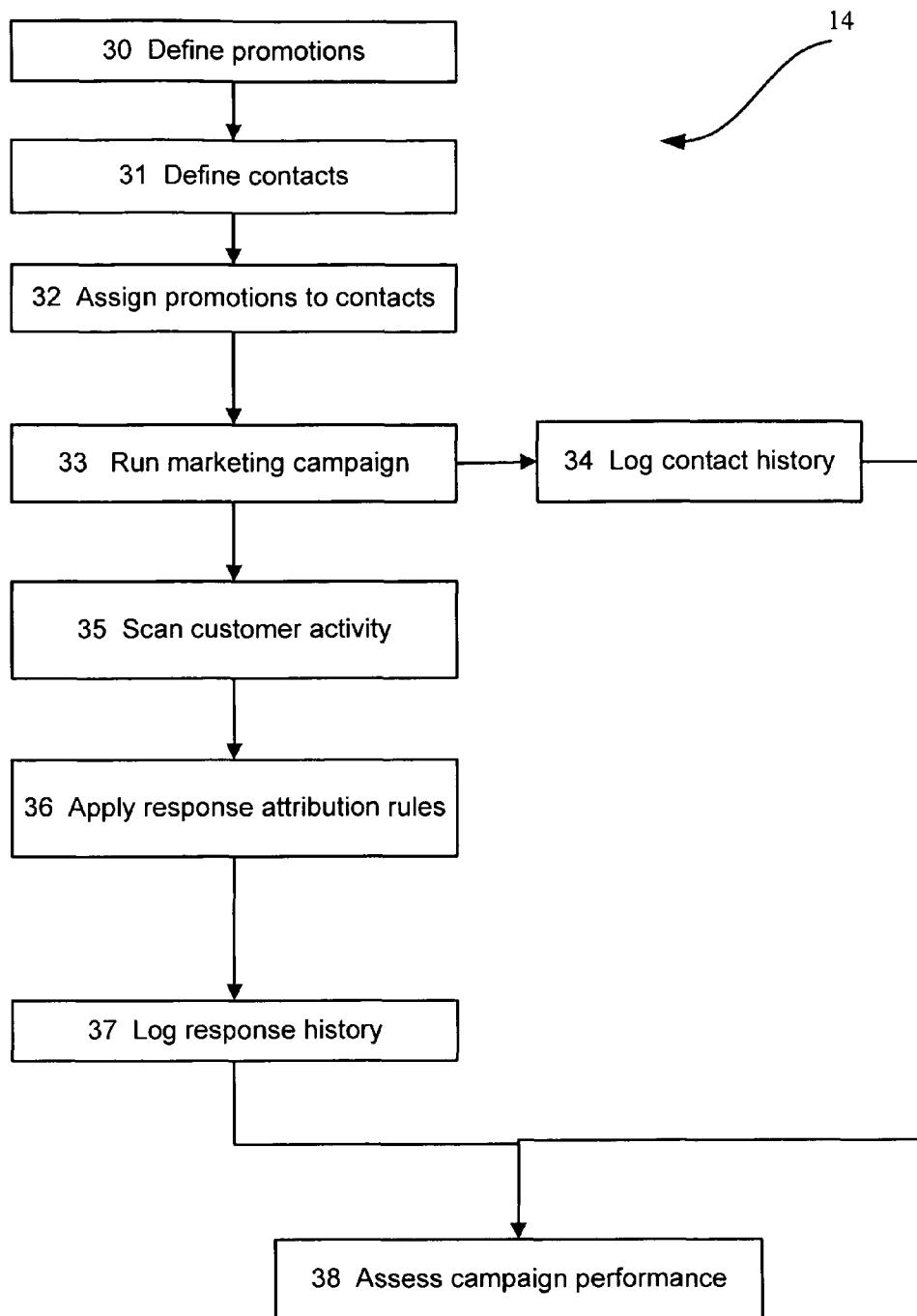


FIG. 2

U.S. Patent

Nov. 20, 2012

Sheet 3 of 26

US 8,315,904 B2

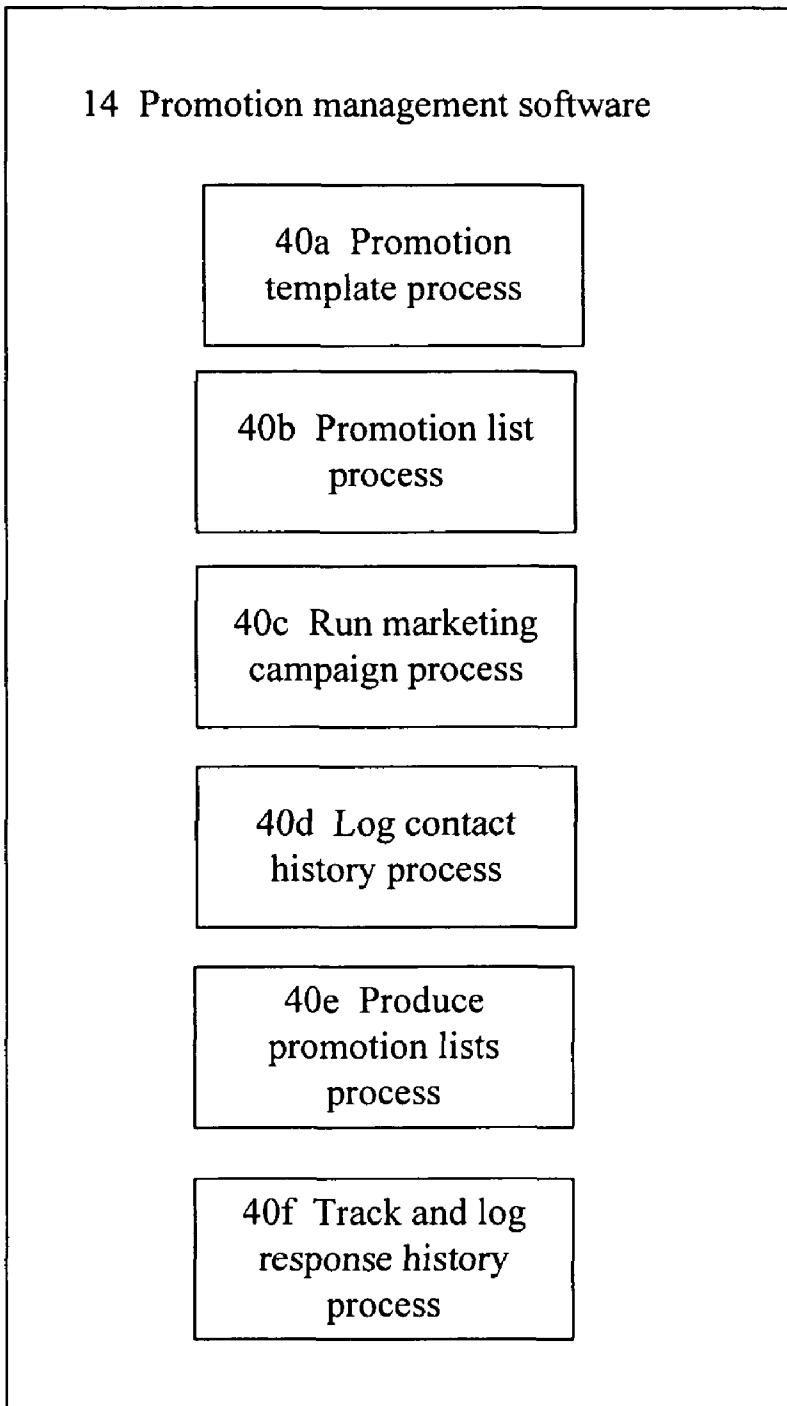


FIG. 3

U.S. Patent

Nov. 20, 2012

Sheet 4 of 26

US 8,315,904 B2

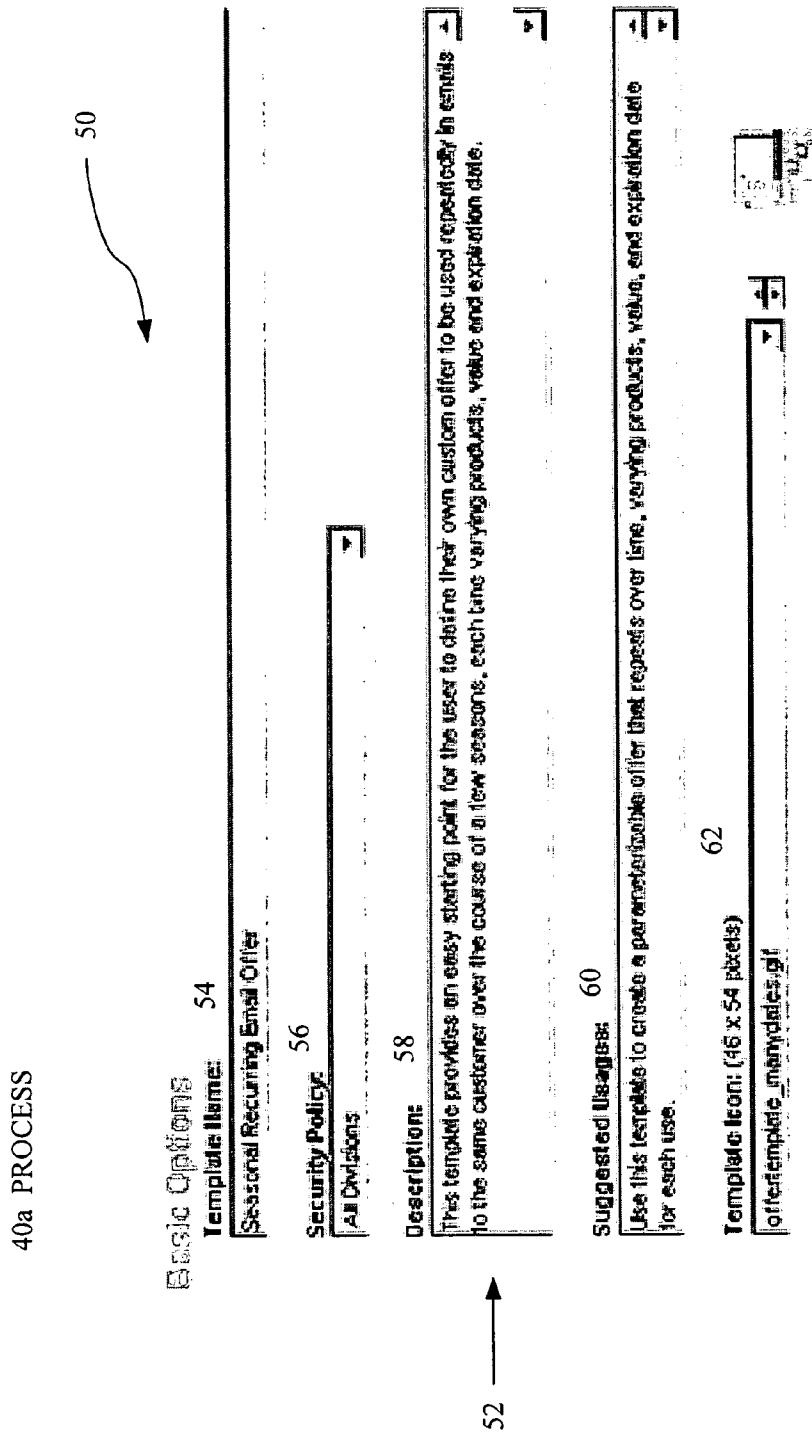


FIG. 4

U.S. Patent

Nov. 20, 2012

Sheet 5 of 26

US 8,315,904 B2

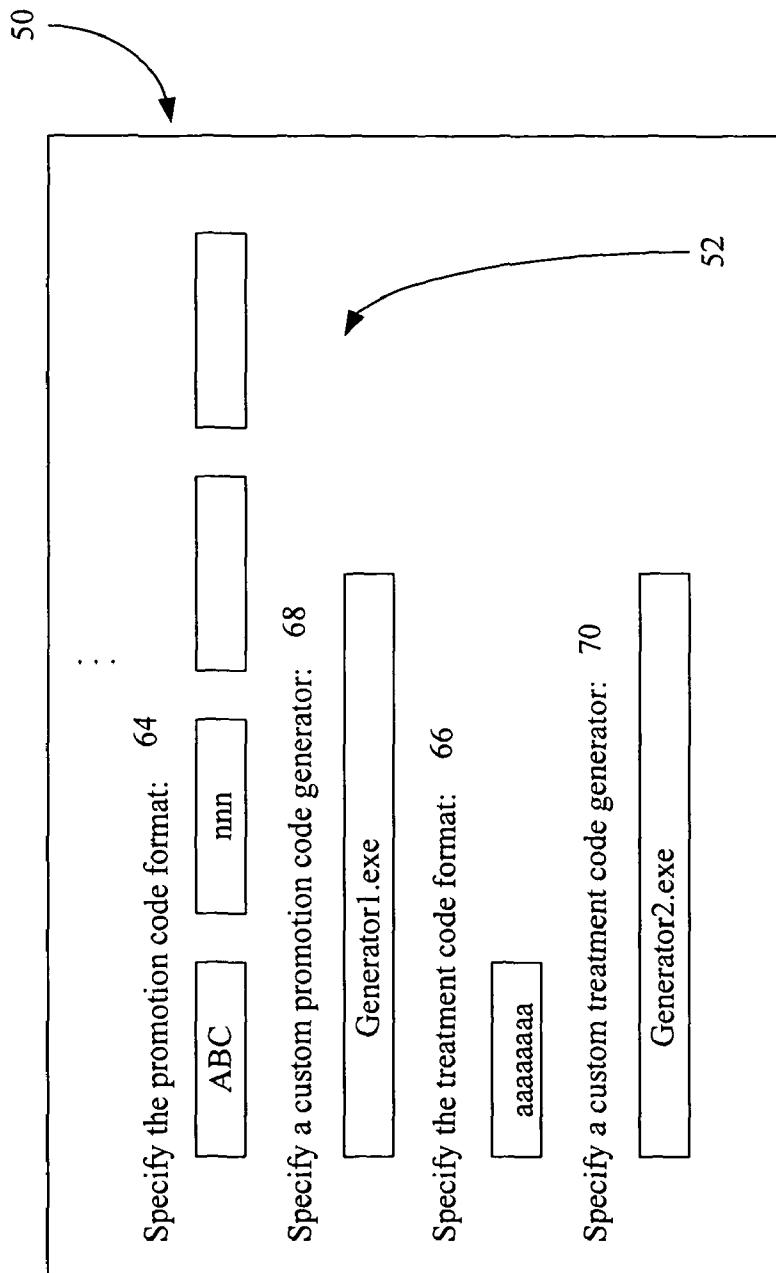


FIG. 5

U.S. Patent

Nov. 20, 2012

Sheet 6 of 26

US 8,315,904 B2

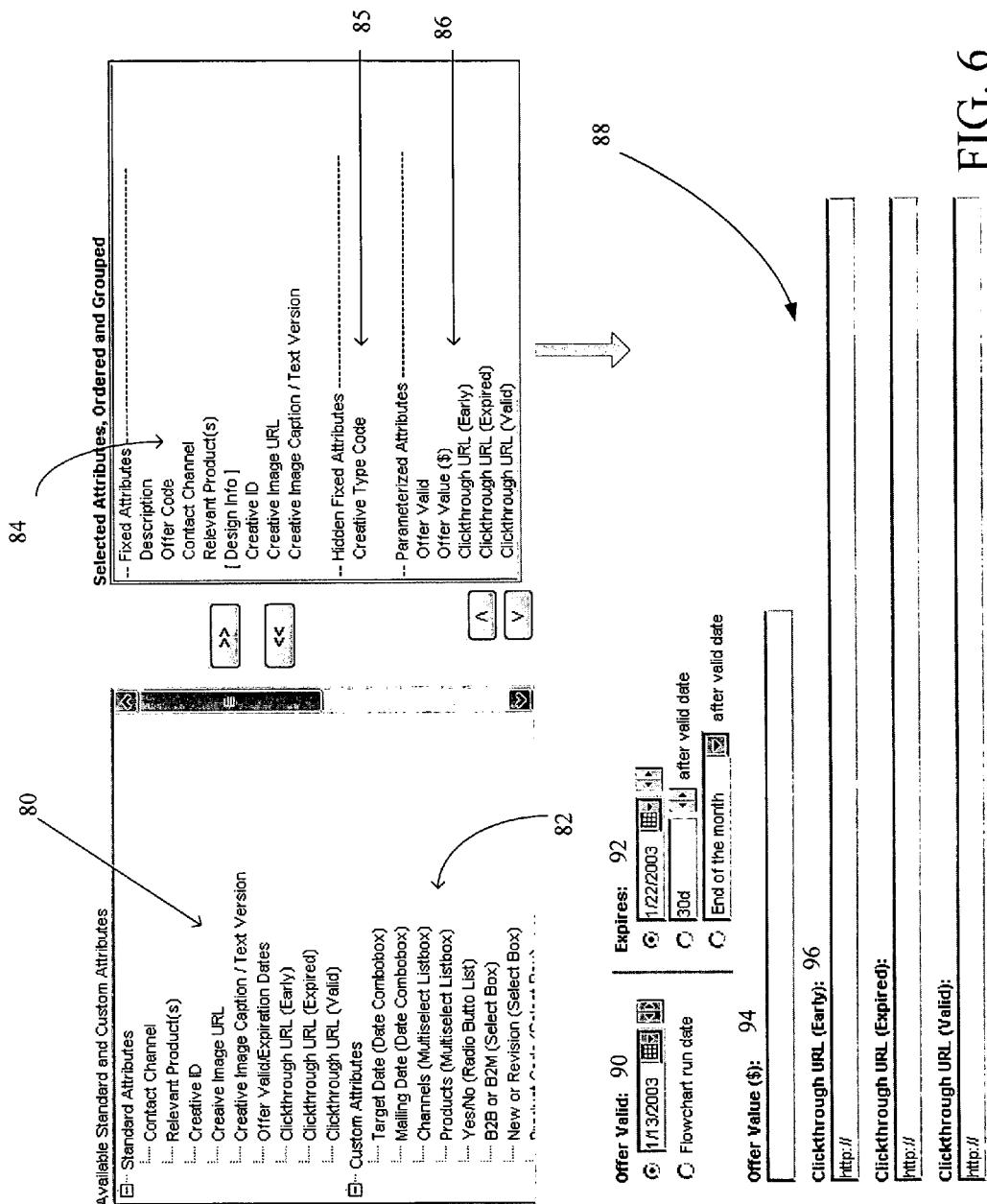


FIG. 6

U.S. Patent

Nov. 20, 2012

Sheet 7 of 26

US 8,315,904 B2

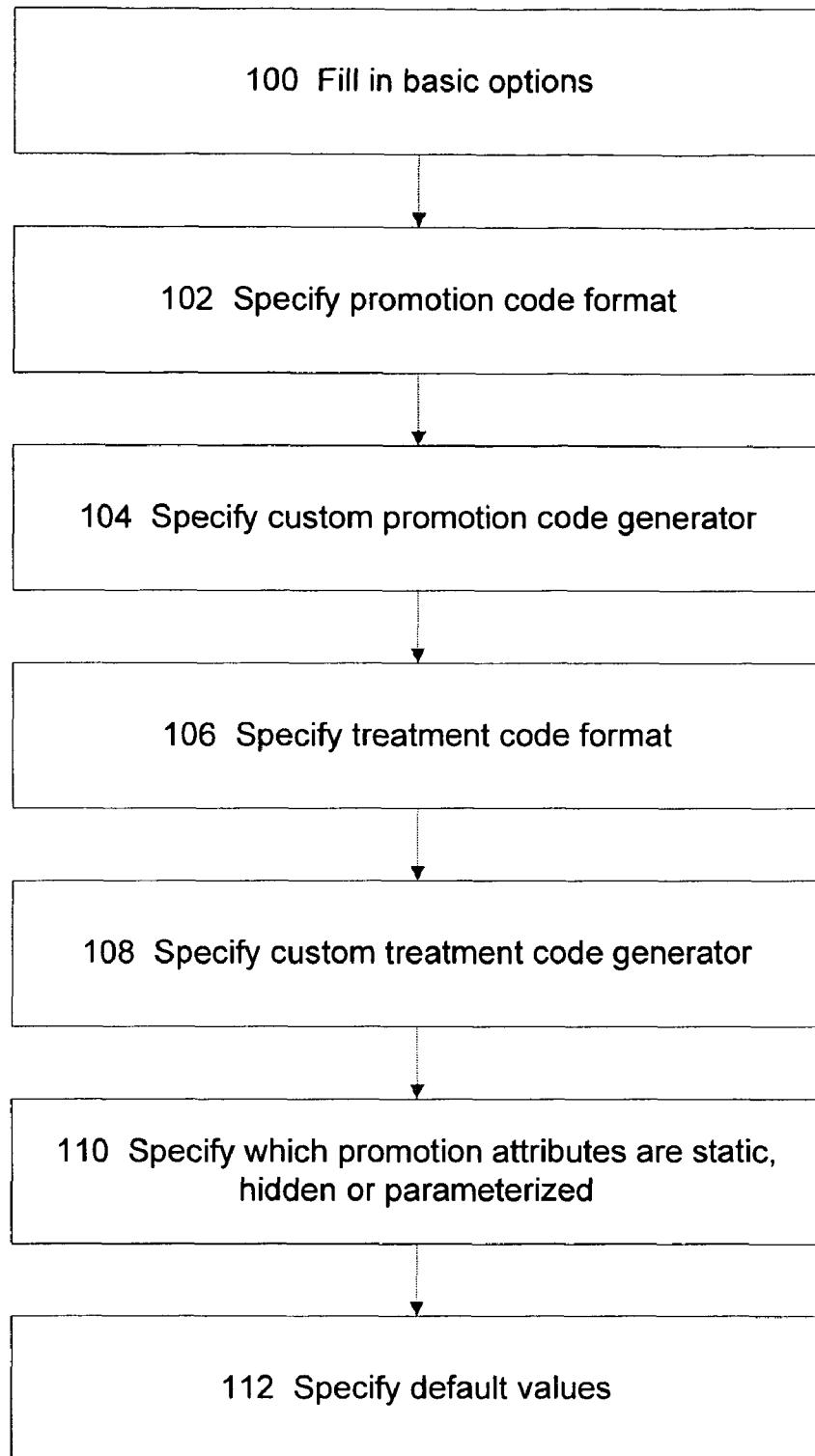


FIG. 7

U.S. Patent

Nov. 20, 2012

Sheet 8 of 26

US 8,315,904 B2

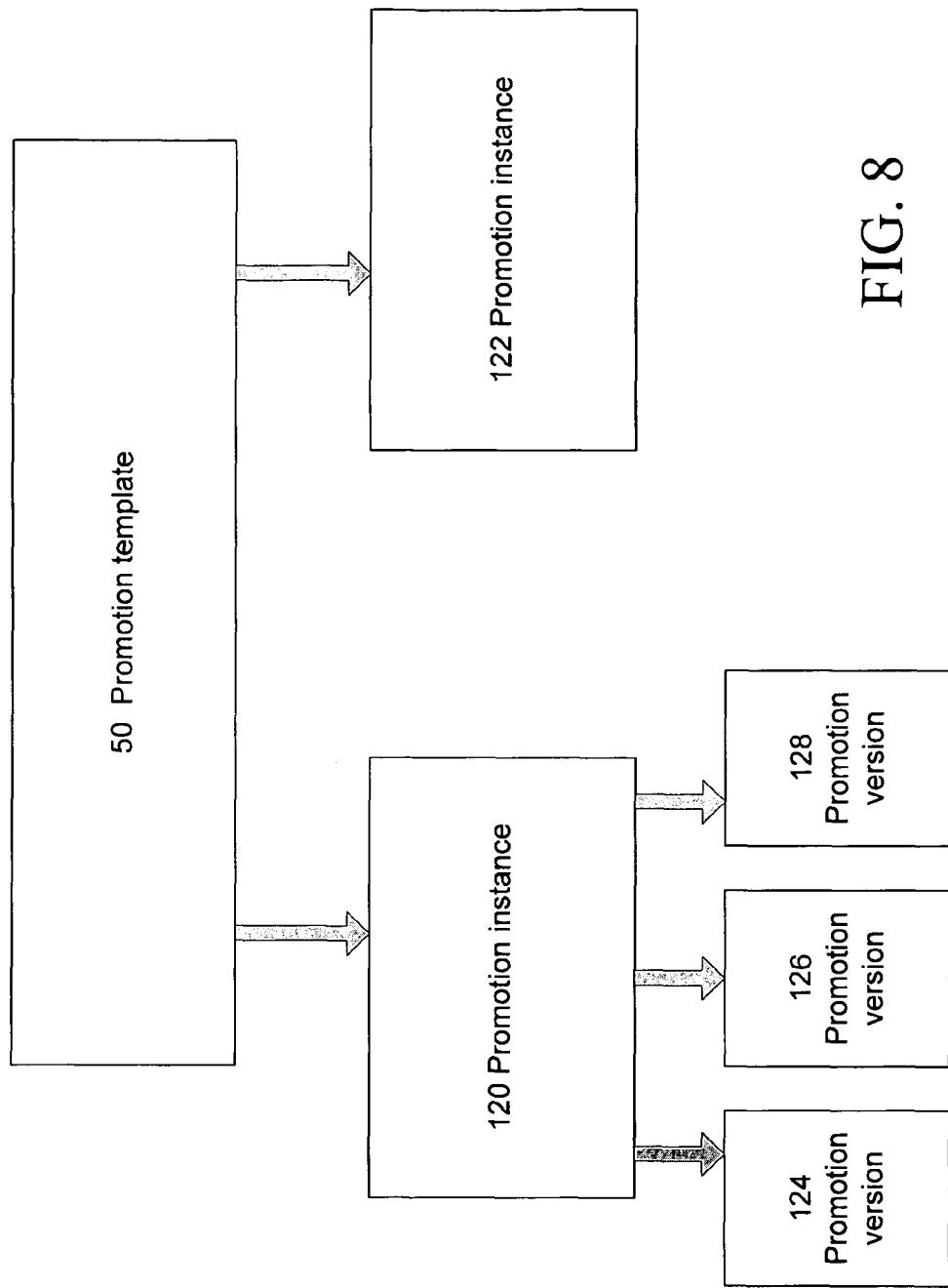


FIG. 8

U.S. Patent

Nov. 20, 2012

Sheet 9 of 26

US 8,315,904 B2

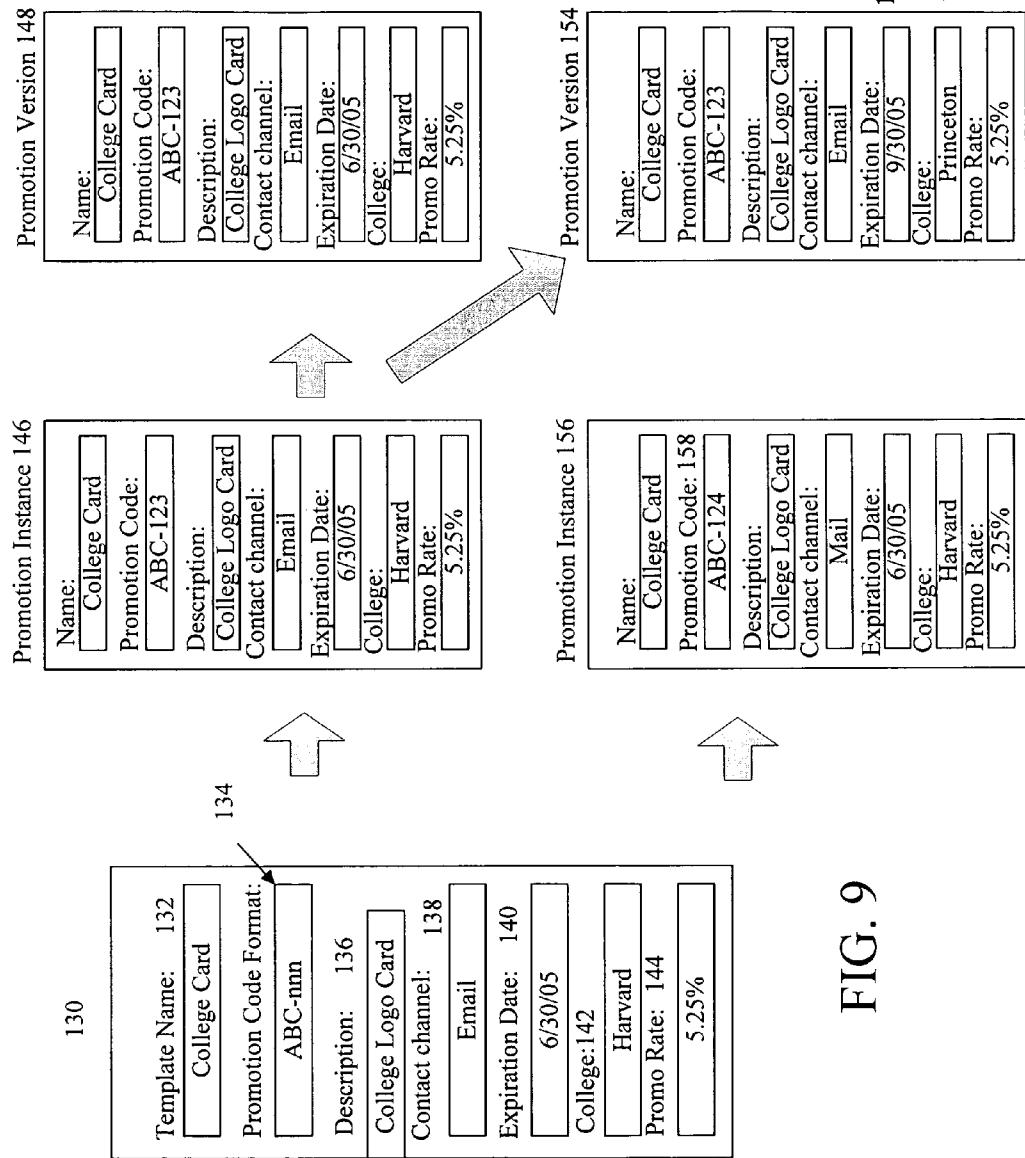


FIG. 9

U.S. Patent

Nov. 20, 2012

Sheet 10 of 26

US 8,315,904 B2

162

Favorite Trips			
CID	From	To	Date
3254673	Boston	New York	39
3254678	Boston	Montreal	86
3254678	New York	London	75
8452353	Atlanta	Chicago	87
8452353	Atlanta	Los Angeles	81
8452353	Atlanta	San Diego	78
8452353	Atlanta	Boston	75
1547653	Chicago	Dallas	93
...

Customer Offers		Offer 5
Offer name	Favorite Destinations	
Offer code	XYZ-789	
Description	Specials on favorite dest.	
Contact channel	Email	
Expiration date	7/30/2005	
From: (departure)	<FavoriteTrips.From>	
To: (destination)	<FavoriteTrips.To>	
Fare	<FavoriteTrips.Fare>	

160

FIG. 10

U.S. Patent

Nov. 20, 2012

Sheet 11 of 26

US 8,315,904 B2

40b Process

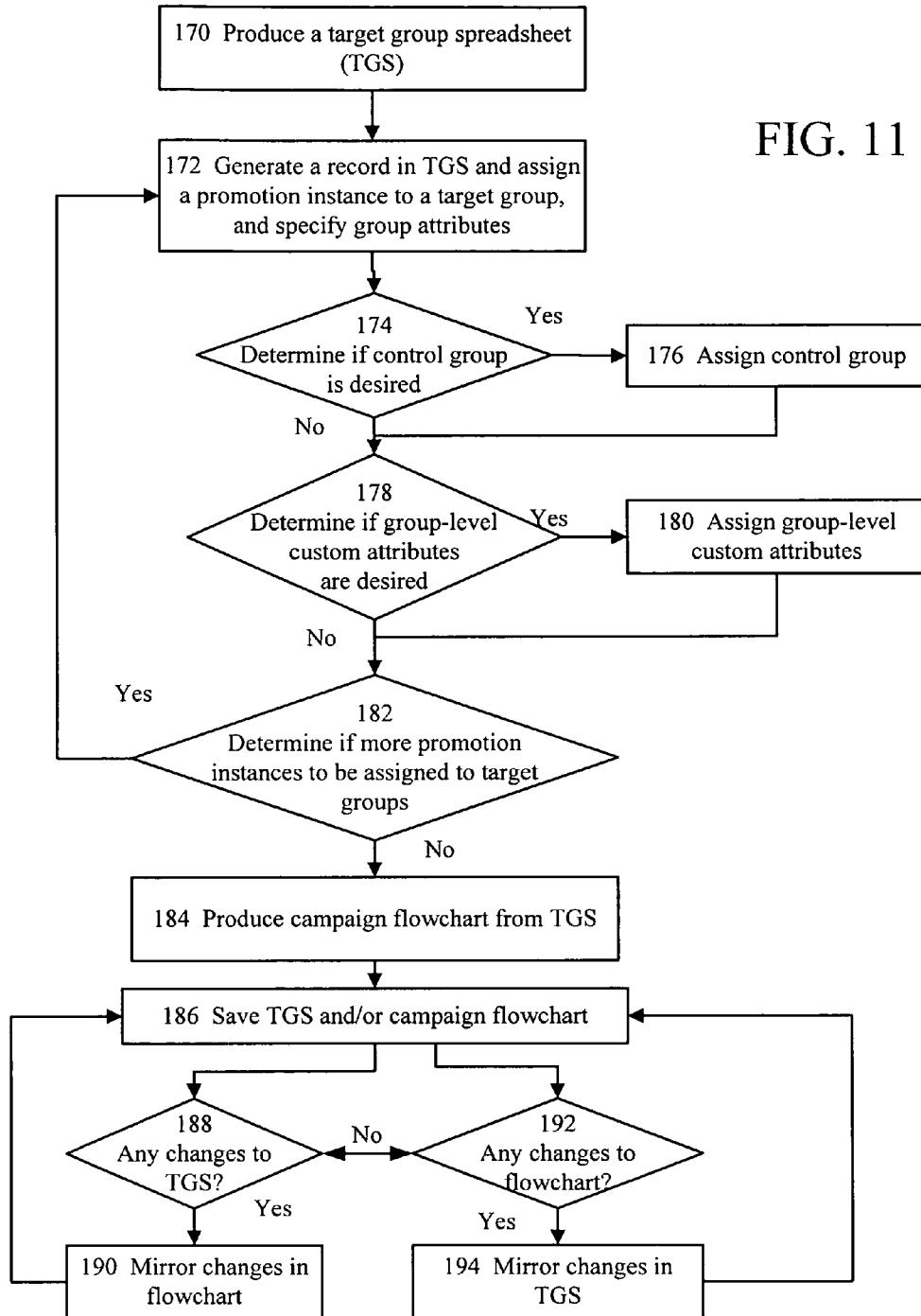


FIG. 11

U.S. Patent

Nov. 20, 2012

Sheet 12 of 26

US 8,315,904 B2

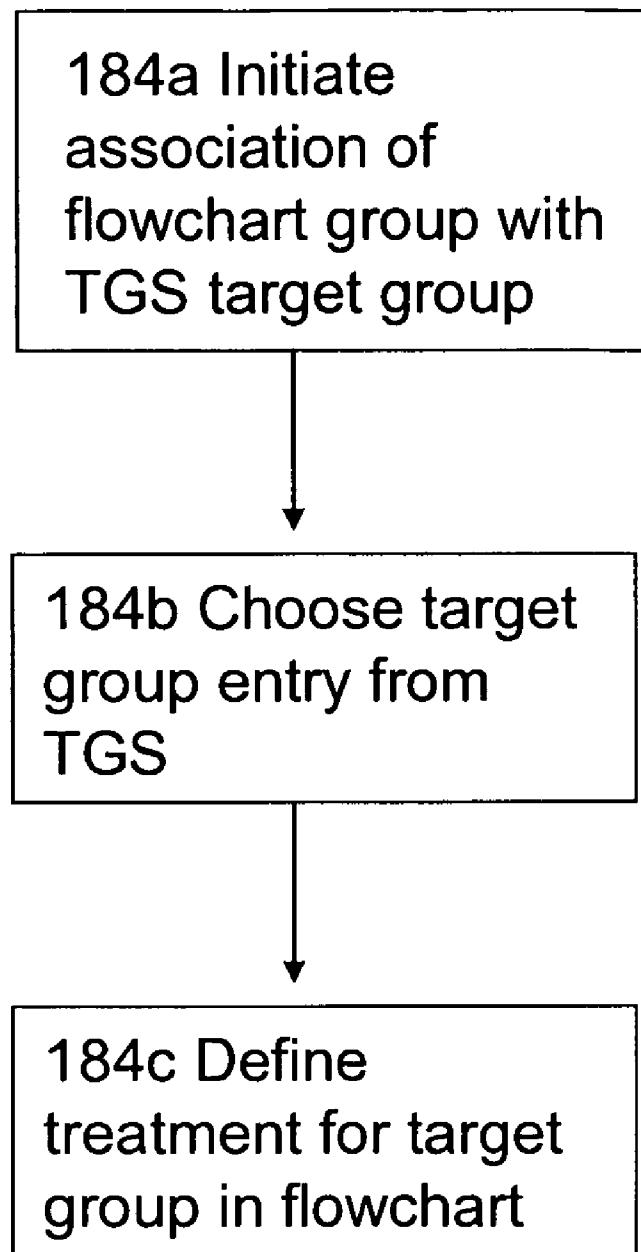


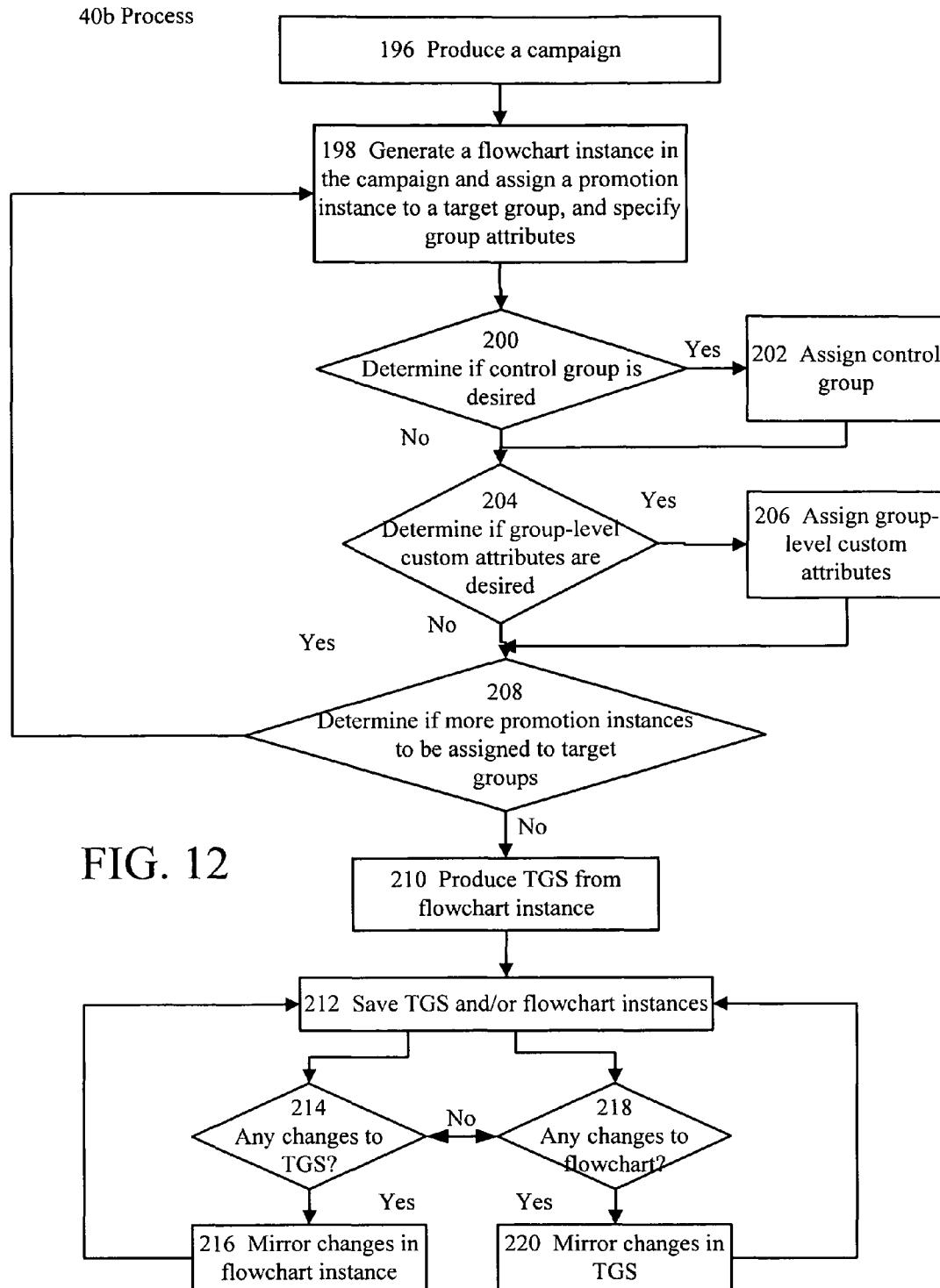
FIG. 11A

U.S. Patent

Nov. 20, 2012

Sheet 13 of 26

US 8,315,904 B2



U.S. Patent

Nov. 20, 2012

Sheet 14 of 26

US 8,315,904 B2

All Campaigns > NorthEast Region > 222

Migration Campaign Cell Attributes | [Edit...](#) | [Mode...](#)

Target Cell Code	Target Cell Name	Control Cell Assigned Code	Offer(s)	Used in Flowchart	Comments
T01-HVHR	HVal_HRisk		Low APR with bal transfer (APR456H); NewsletterTabularOffers (STRAG03GH); SendToAFriend_FreeShipping (SILMA000AH)	Multi-channel	
T01-HVMR	HVal_MedRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T01-HVLR	HVal_LoRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T02-MVHR	MedVal_HRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T02-MVMR	MedVal_MedRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T02-MVLR	MedVal_LoRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T03-LVHR	LoVal_HRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T03-LVMR	LoVal_MedRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T03-LVLR	LoVal_LoRisk		Low APR with bal transfer (APR456H)	Multi-channel	
T04-LVMR	LoVal_MedRisk		Low APR with bal transfer (APR456H)		
T04-LVLR	LoVal_LoRisk		Low APR with bal transfer (APR456H)		
T05-LVLR	LoVal_LoRisk		Low APR with bal transfer (APR456H)	Multi-channel	



All Campaigns > NorthEast Region > 220

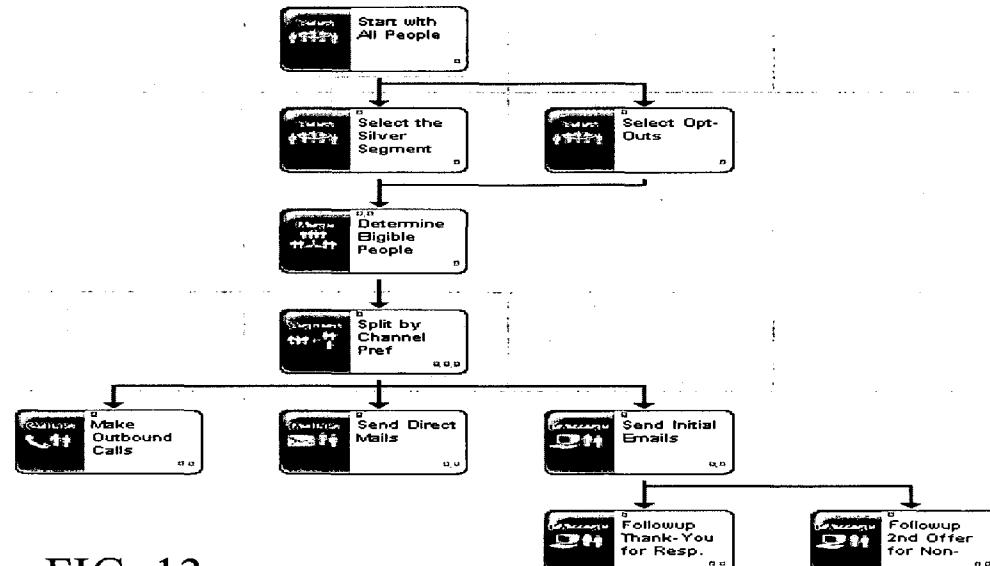
Migration Campaign Multi-channel | [Edit...](#) | [Run...](#)

FIG. 13

U.S. Patent

Nov. 20, 2012

Sheet 15 of 26

US 8,315,904 B2

40c PROCESS

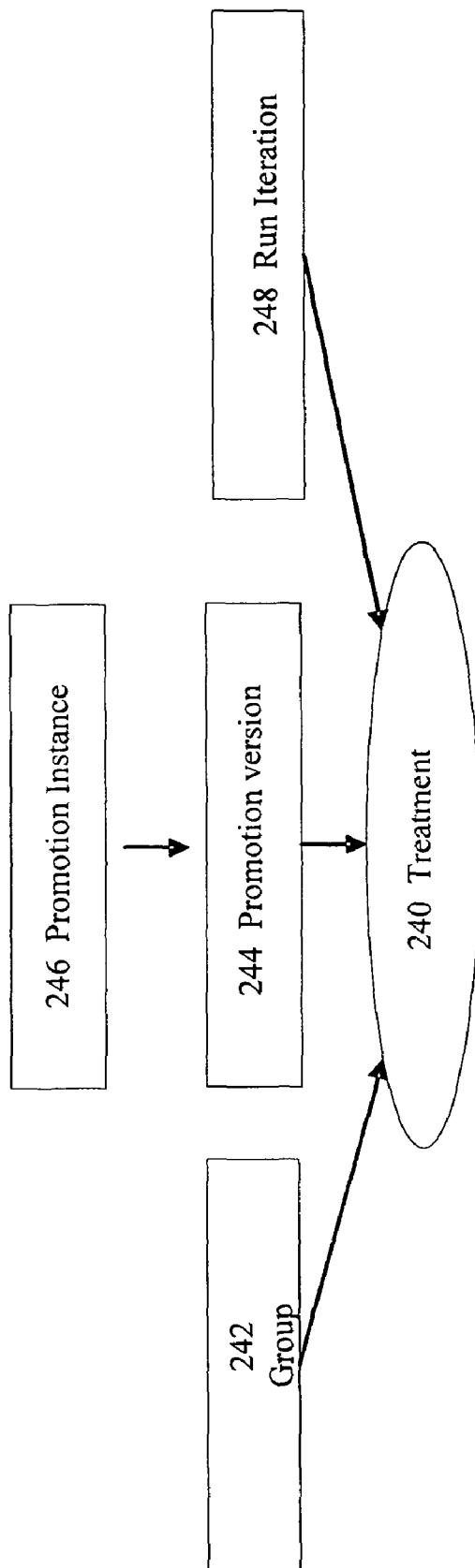


FIG. 14

U.S. Patent

Nov. 20, 2012

Sheet 16 of 26

US 8,315,904 B2

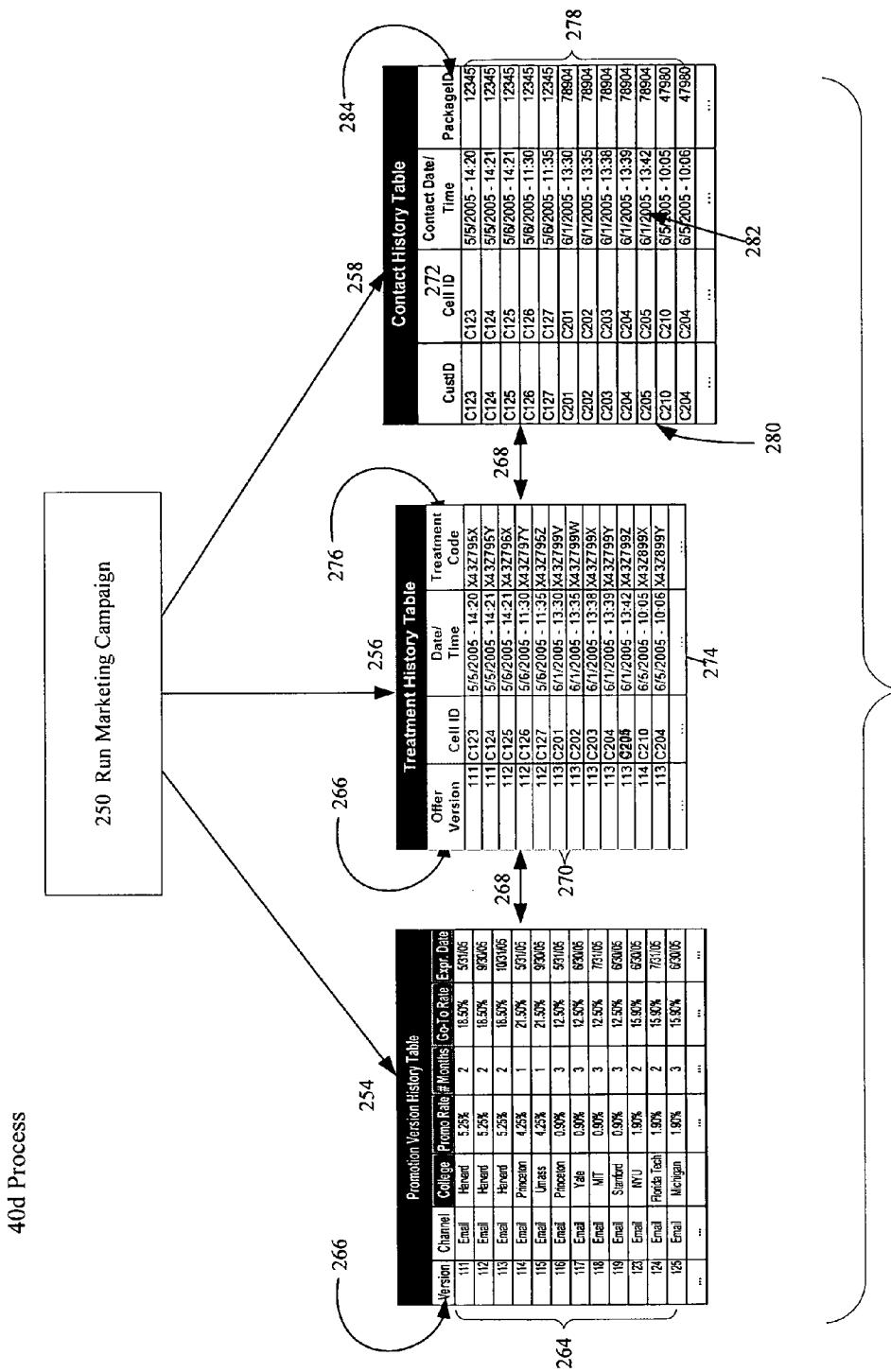


FIG. 15

U.S. Patent

Nov. 20, 2012

Sheet 17 of 26

US 8,315,904 B2

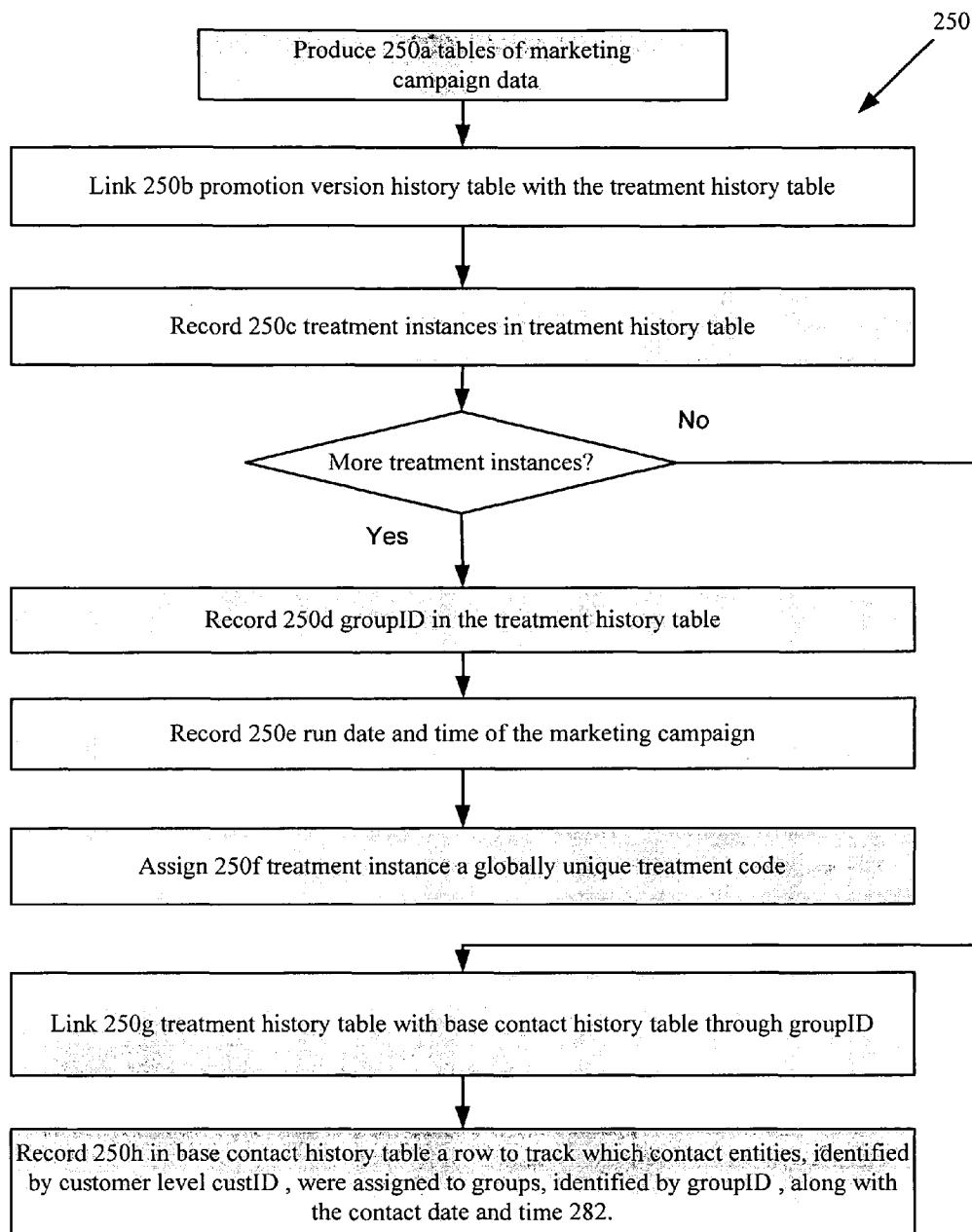


FIG. 15A

U.S. Patent

Nov. 20, 2012

Sheet 18 of 26

US 8,315,904 B2

300

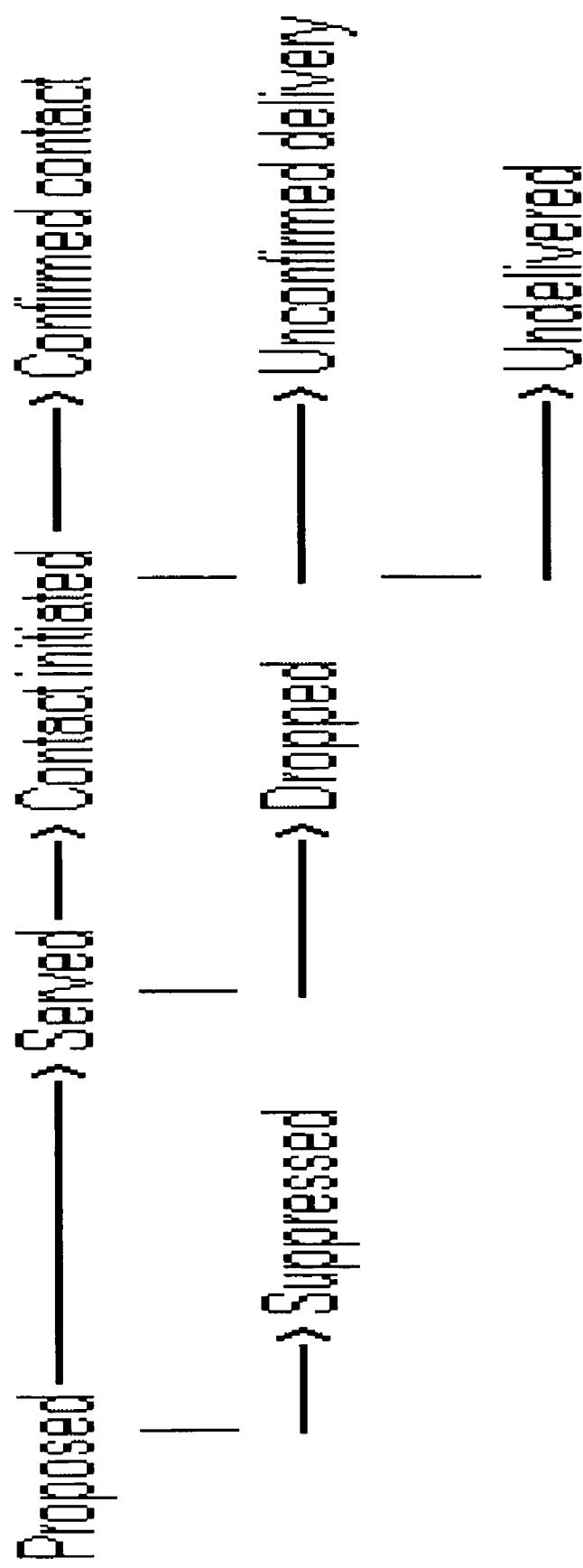


FIG. 15B

U.S. Patent

Nov. 20, 2012

Sheet 19 of 26

US 8,315,904 B2

40e PROCESS

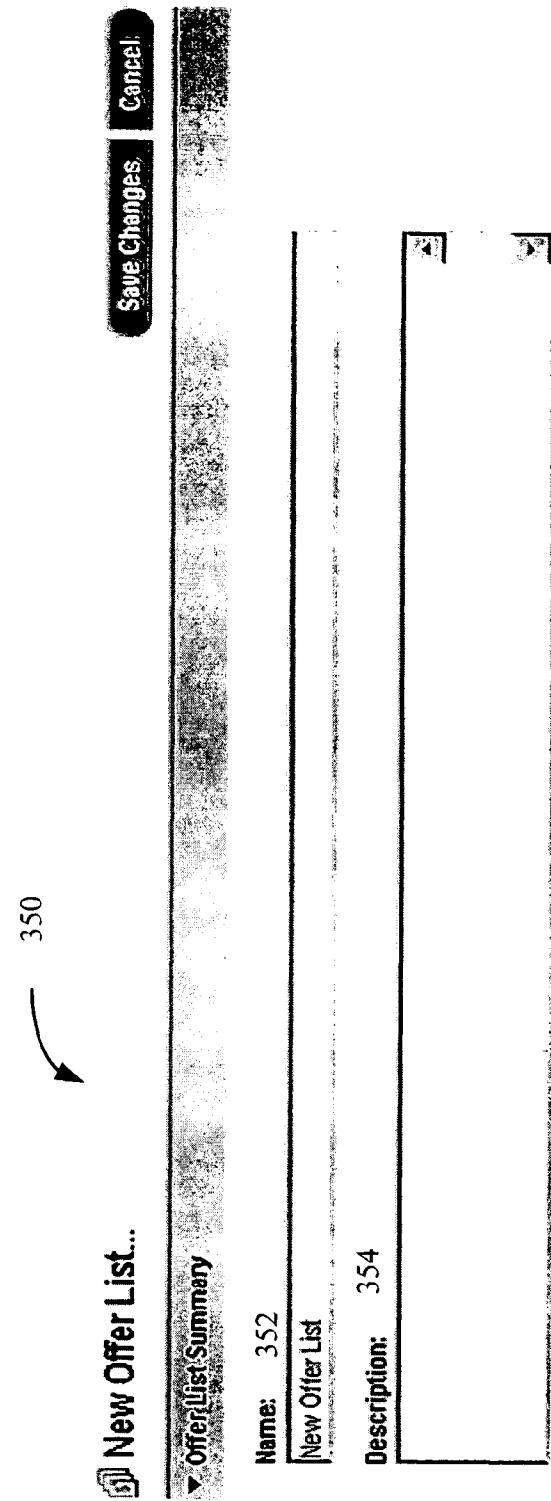


FIG. 16

U.S. Patent

Nov. 20, 2012

Sheet 20 of 26

US 8,315,904 B2

356 FOLDER TREE VIEW

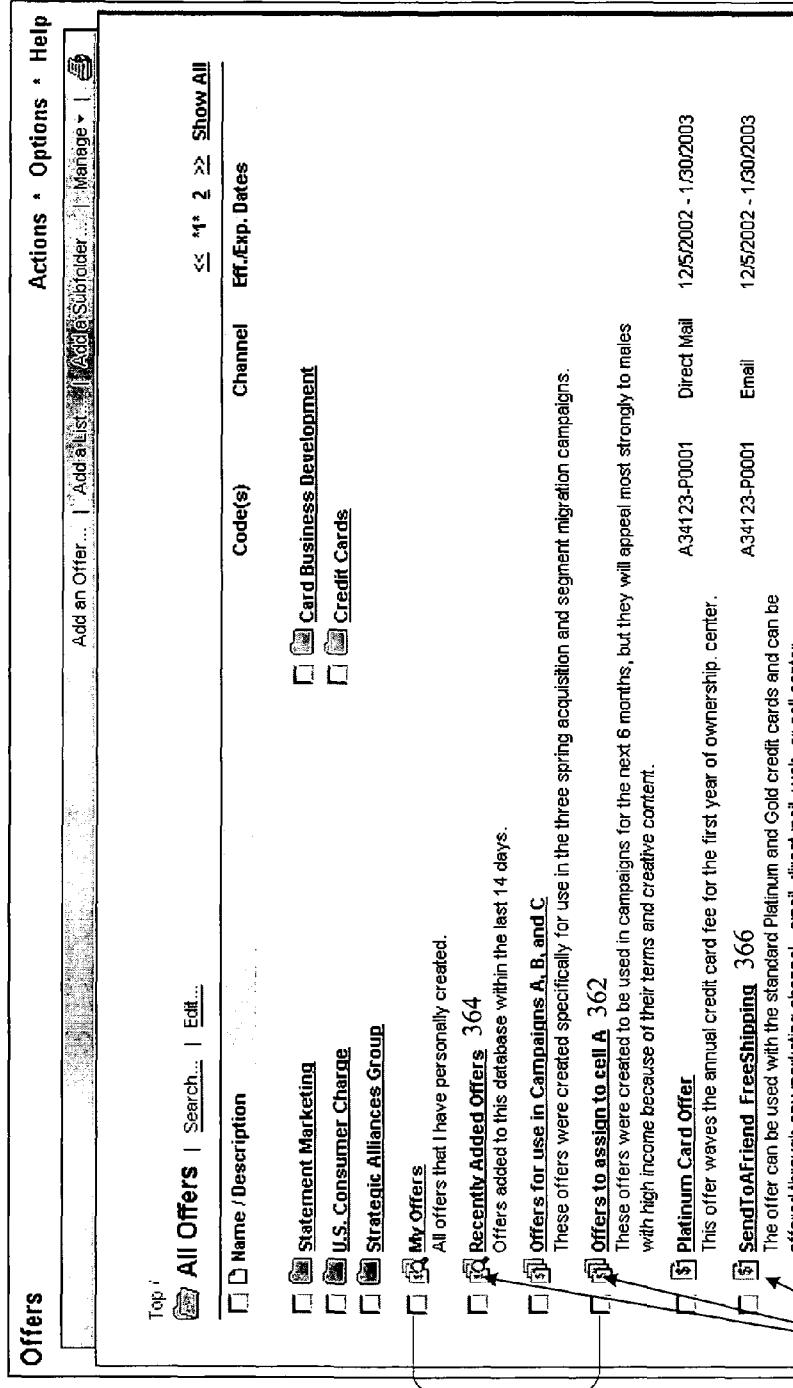


FIG. 17

360
358 PROMOTION LISTS

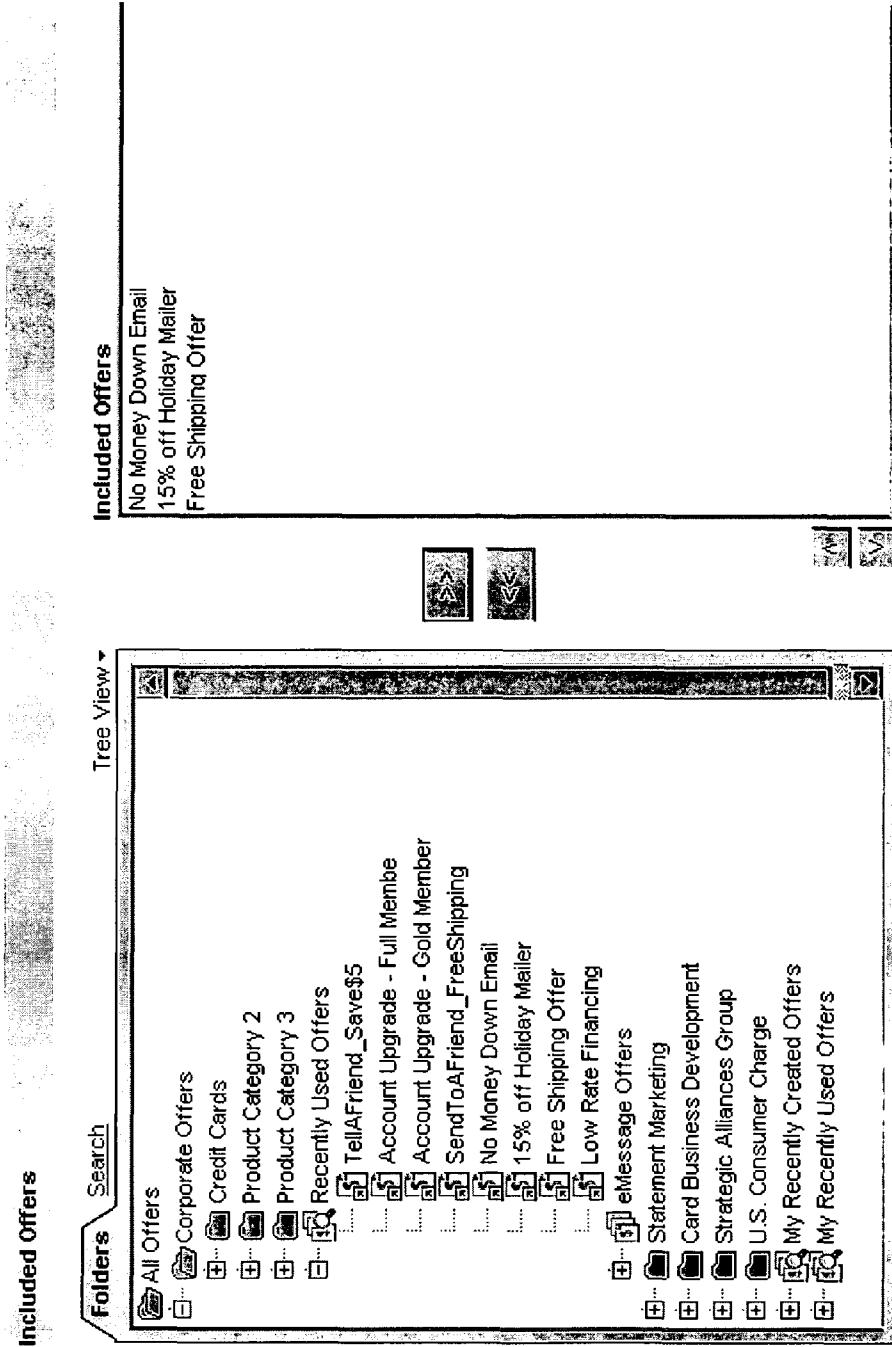
U.S. Patent

Nov. 20, 2012

Sheet 21 of 26

US 8,315,904 B2

356 FOLDER TREE VIEW



37

FIG. 18

U.S. Patent

Nov. 20, 2012

Sheet 22 of 26

US 8,315,904 B2

372 QUERY

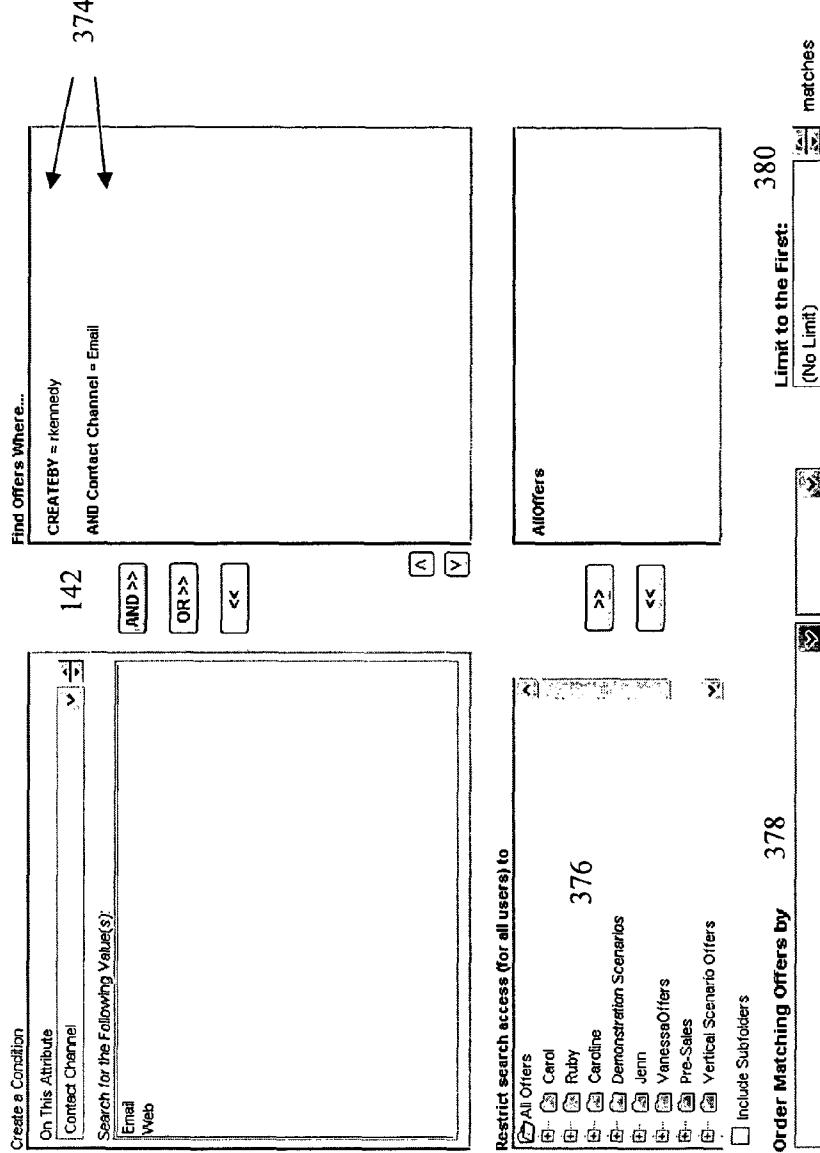


FIG. 19

U.S. Patent

Nov. 20, 2012

Sheet 23 of 26

US 8,315,904 B2

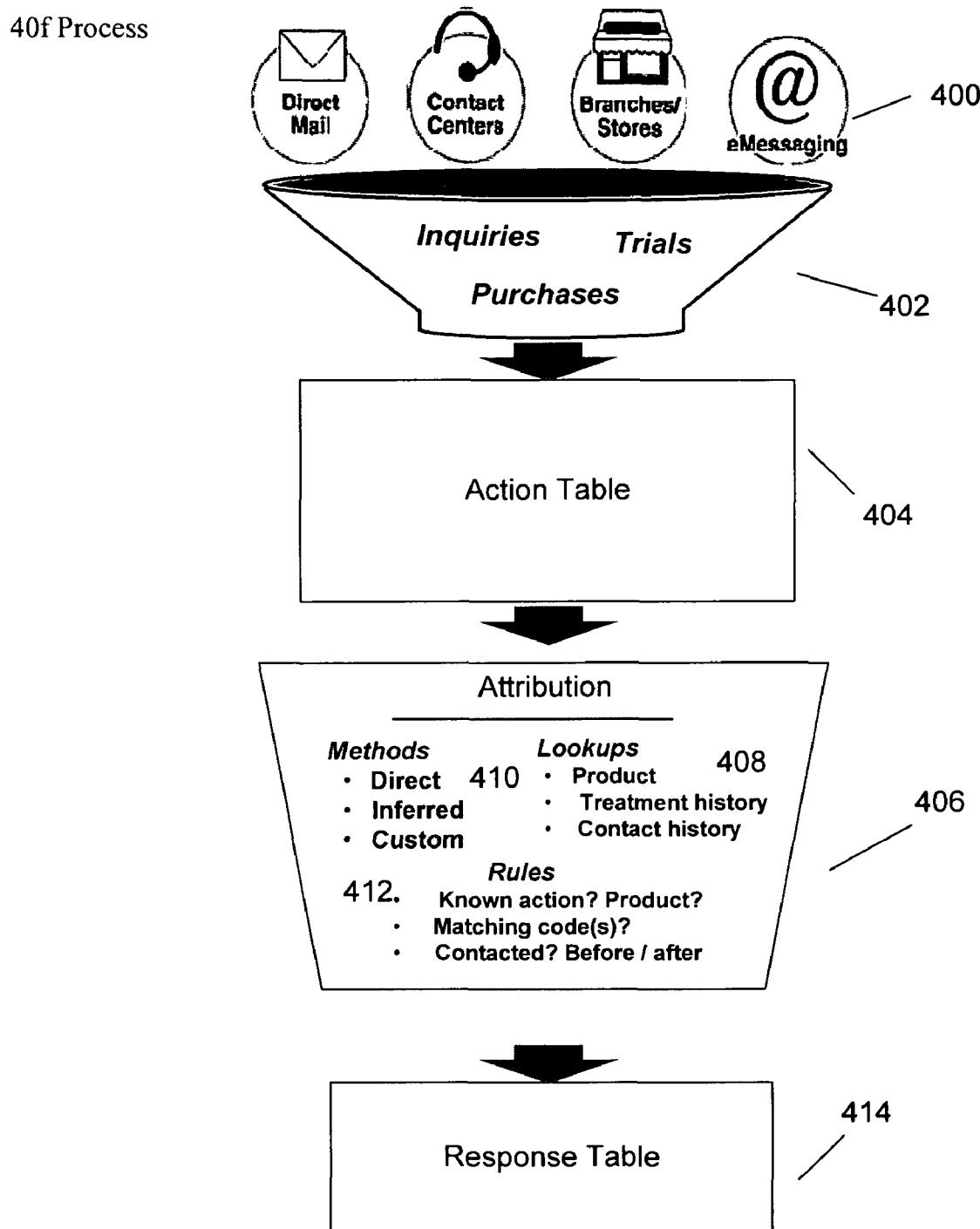


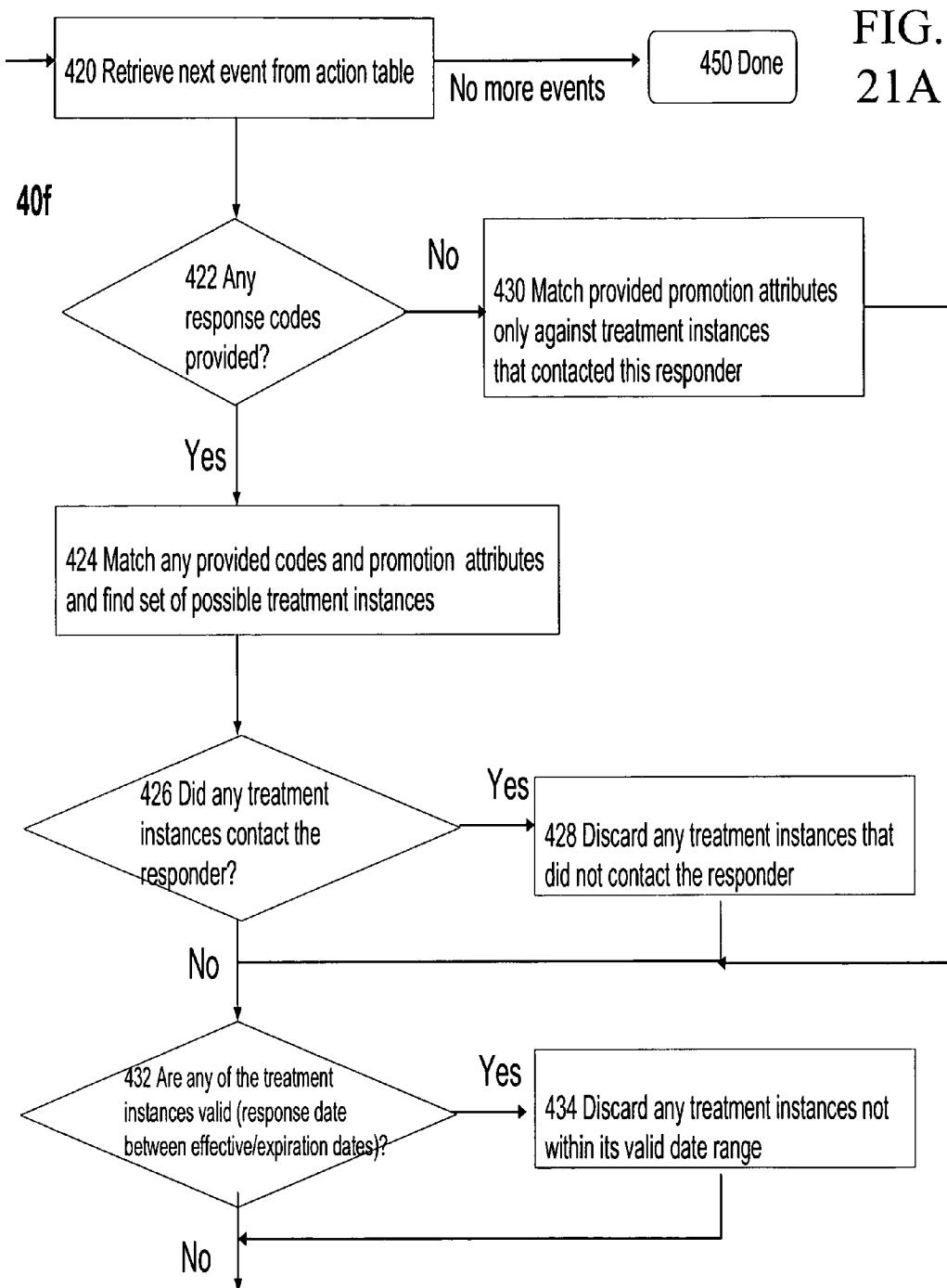
FIG. 20

U.S. Patent

Nov. 20, 2012

Sheet 24 of 26

US 8,315,904 B2



U.S. Patent

Nov. 20, 2012

Sheet 25 of 26

US 8,315,904 B2

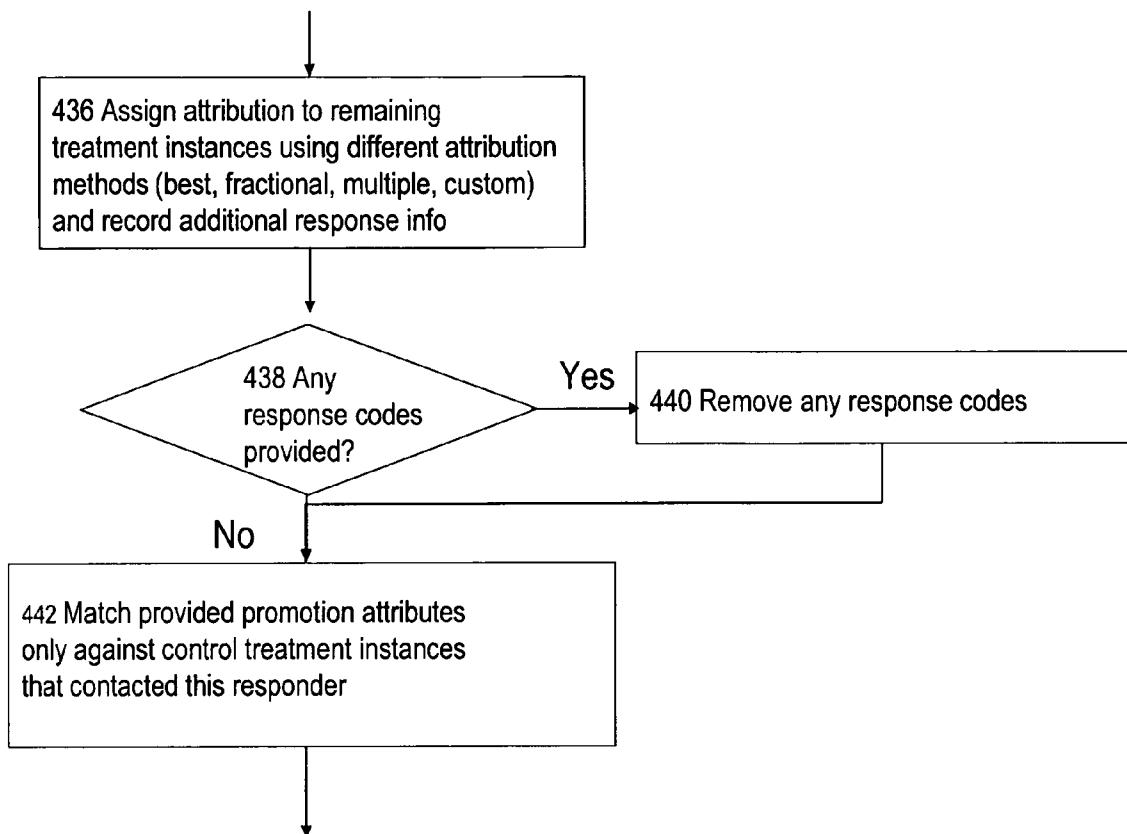


FIG. 21B

U.S. Patent

Nov. 20, 2012

Sheet 26 of 26

US 8,315,904 B2

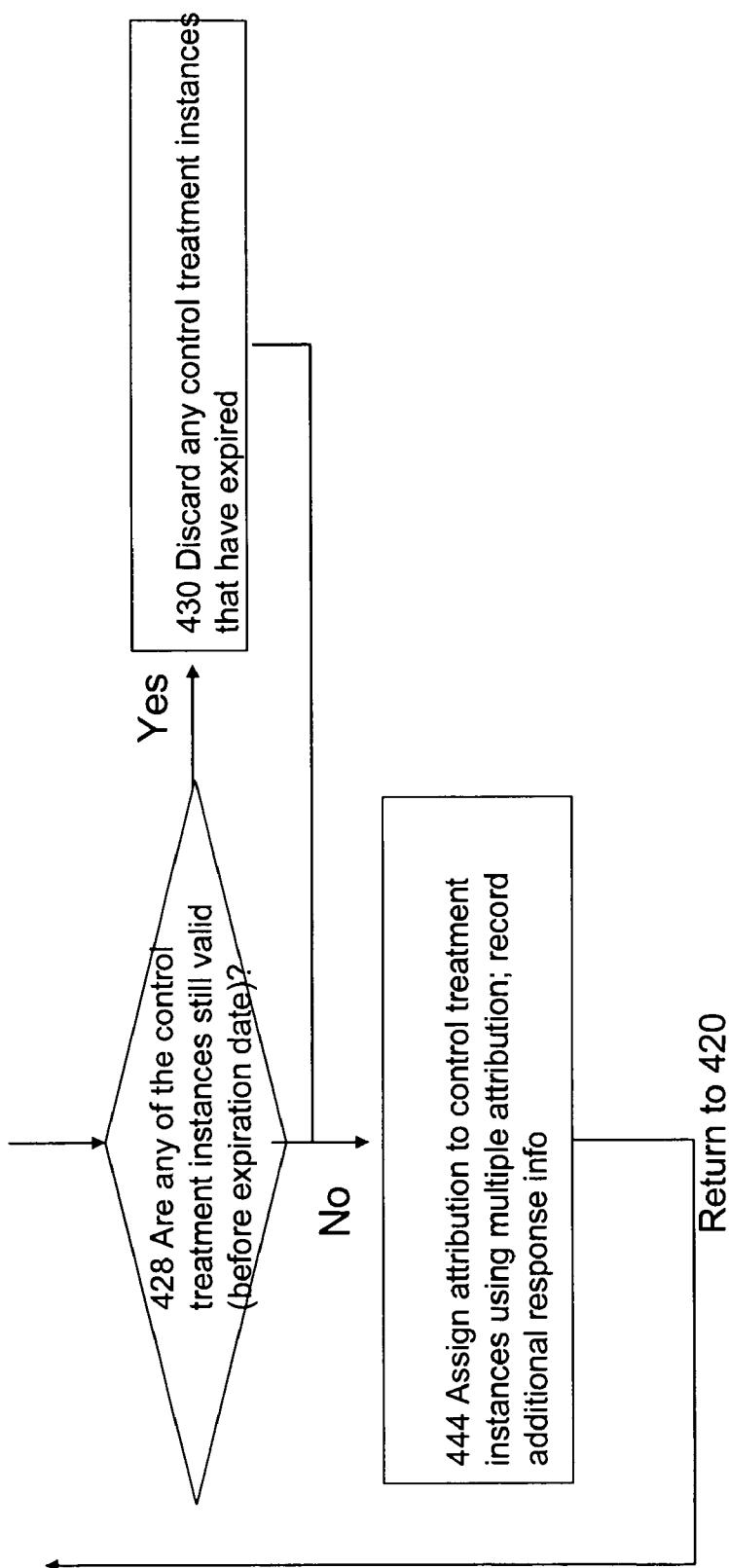


FIG. 21C

US 8,315,904 B2

1**ORGANIZATION FOR PROMOTION
MANAGEMENT****BACKGROUND**

Promotions are specific communications made from one entity to another entity. The challenges in managing promotions can arise out of the various situations in which promotions are used. For example, marketers in the financial services, travel and retail industries may each have specialized considerations when communicating with potential or existing customers. Each industry may have different types of promotions directed to communicate different types of products or services, resulting in an infinite variety of possible attributes. All marketers produce a potentially large variety of promotions and assign them to a potentially large variety of contacts. These promotions are logged and tracked for future analysis to improve the promotions.

SUMMARY

According to an aspect of the present invention, a computer implemented method for producing a promotion list for a promotion management campaign, the method including assigning one or more promotion instances to the promotions list, and storing the promotion list in an electronic medium.

According to another aspect of the present invention, a computer program product residing on a computer readable medium for producing a promotion list for a promotion management campaign, includes instructions for causing a computer to assign one or more promotion instances to the promotions list, and store the promotion list in an electronic medium.

According to another aspect of the present invention, an apparatus includes a processor and memory executing a computer program product for producing a promotion list for a promotion management campaign. The computer program product including instructions to assign one or more promotion instances to the promotions list; and store the promotion list in an electronic medium.

Implementations can include one or more of the following. Assigning the one or more promotions to the promotions list includes querying promotion instances to find one or more matching promotion instances having matching attributes to one or more of the promotion attributes, and at runtime, adding the matching promotion instances to the promotions list. Assigning the one or more promotions to the promotions list includes displaying the one or more promotion instances stored in the electronic medium in a searchable and browsable view, and adding from the searchable and browsable view, one or more promotion instances to the promotions list. Limiting the matching promotion instances to an N number of promotions, wherein N is an integer. Sorting the matching promotion instances in one of an ascending or descending order based on one or more promotion attributes. Filtering the matching promotion instances to the matching promotion instances located in one or more folders of the electronic medium. Assigning to the promotion list a security policy governing which users are allowed to interact with the promotion list. The promotion list attributes include a promotion list name, an icon, and a promotion list description. The query includes clauses including one or more of including a function, a variable, a promotion attribute connected by Boolean operators to match to one or more attributes of one or more promotions. One or more promotions include a first promo-

2

tion having at least one or more different promotion attributes or different promotion attribute values than a second promotion.

DESCRIPTION OF DRAWINGS

FIG. 1 is a block diagram of a computer system executing promotion management software.

FIG. 2 is a flowchart depicting the promotion management lifecycle.

FIG. 3 is a block diagram of the promotion management software.

FIGS. 4 and 5 are diagrams depicting a promotion template.

FIG. 6 is a diagram depicting standard attributes and custom attributes of a promotion template.

FIG. 7 is a flowchart for producing a promotion template.

FIGS. 8 and 9 are diagrams depicting promotion instances and promotion versions.

FIG. 10 is a diagram depicting data-driven parameterization of promotions.

FIGS. 11, 11A, and 12 are flowcharts depicting the target group management process.

FIG. 13 is a diagram depicting a target group spreadsheet and a campaign flowchart.

FIG. 14 is a flowchart depicting producing treatment instances.

FIGS. 15 and 15A are a diagram and flow chart depicting contact history and a contact history process.

FIG. 15B is a tree depicting the hierarchy of contact history.

FIG. 16 is a diagram depicting promotion list attributes.

FIG. 17 is a diagram depicting a folder tree view of the promotion list process.

FIG. 18 is a diagram depicting the selection of offers in a static promotion list.

FIG. 19 is a diagram depicting a query in a dynamic promotion list.

FIG. 20 is a general flowchart depicting the high-level response tracking process.

FIG. 21A to 21C are flowcharts depicting the steps in tracking direct and inferred responses and the inferred responses for control groups.

DETAILED DESCRIPTION

Referring to FIG. 1, a network computer system 10 includes a server 12 executing software for promotion management 14, connected through a network 16, to clients 18a to 18i executing respective client instances 20a to 20i of promotion management software 14. As used herein, the terms "promotion(s)" and "offer(s)" are used interchangeably, but are limited to refer to specific marketing communications that are typically provided in a marketing campaign. Although the network computer system 10 is shown including the server 12 running promotion management software server process 14 and client systems 18a to 18i running the client promotion management software 20a to 20i, it is understood that the server software 14 and the client promotion management software 20a to 20i can run on the same machines, e.g., the server 12. The client systems are used to interface with promotion management software instances 20a to 20i to produce and manage promotions. The server 12 is connected to an application database 22 that stores information required by the promotion management software 14 to operate (e.g., system table data) and a marketing database 24 that stores the company marketing (e.g., customer or prospect) and related

US 8,315,904 B2

3

data and can be used to drive marketing campaigns. The application and marketing databases 22 and 24 may run on the same machine as server 12 or different machines. Databases 22 and 24 may be in different databases, on different machines, or they may be separate schemas in the same database on the same machine. The clients 18a-18i are connected to the server 12 by the network 16.

Referring to FIG. 2, promotion management software 14 is shown. A user defines one or more promotions 30 that the user wishes to use in a marketing campaign. The user defines the contacts 31, which can be persons or entities (e.g., households, businesses, accounts of individuals), or both, for the marketing campaign. The promotions are assigned to the contacts 32. The promotion management software 14 (from FIG. 1) runs the marketing campaign 33. The promotion management software 14 logs the contact history of the marketing campaign 34, to maintain a record of which contacts received what promotions. Logging of the contact history 34 also facilitates assessing the campaign's performance, as discussed below.

After the promotion management software 14 runs the marketing campaign 33, the software scans relevant customer activity 35 and applies response attribution rules 36 to the customer activity. The response attribution rules, as will be discussed in detail below, determine which promotions receive credit for customer activity. The promotion management software 14 logs the response history 37. The logged contact history 34 and logged response history 37 data are used to assess the marketing campaign's performance 38.

Referring to FIG. 3, the promotion management software instances 20a to 20i (in FIG. 1) include software processes 40a to 40f such as, for example, processes that produce promotion templates, promotion lists, and processes that manage contact histories and target groups. As used herein, groups refer to collections of contacts, typically individuals or other entities, whereas target groups refer to groups of entities that are intended recipients for specific marketing communications, e.g., promotions, and control groups refer to groups of entities that are typically randomly sampled from the target groups and withheld from receiving the marketing communications to evaluate the effectiveness of the marketing communications, as will be further discussed below. Control groups are used by marketers for comparison purposes against the target group. This can be a hold-out control group which receives no communication, or it may be a test/challenger group compared to a base/champion group where the test/challenger group receives an alternate "test" promotion. This is also known in marketing as A/B splits. When control groups are used for champion/challenger comparisons, there can be many challenger groups associated with the champion group. In either the hold-out control group or the champion/challenger scenario ("test group scenario"), one or more comparisons are made between promotions assigned to the control group vs. the target group.

Process 40a produces promotion templates and promotion instances. Process 40b manages the design of a marketing campaign and assignment of promotion instances to groups of contact entities. Process 40c executes the marketing campaign. Process 40d manages the contacts of a marketing campaign and logs the history of the promotions made to groups of contacts, including members of control groups. Process 40e produces contact lists (not including control groups) for the promotions that aid the user in managing and optimizing the marketing campaign. Process 40f logs and tracks the responses from contact entities and controls.

Referring now to FIGS. 4 and 5, process 40a included in the promotion management software produces a promotion

4

template 50 that includes a set of fields that are "basic options" 52. The basic options attribute fields 52 are pre-assigned by the process 40a to define a promotion to the promotion management software. Values of the basic options attribute fields 52 are assigned to the promotion template 50 by the user (typically an application administrator). Each time the user produces a promotion template 50, the user chooses and populates attribute fields for basic options 52.

The basic options include a template name field 54, a security policy field 56, a description field 58, suggested usages fields 60, and an icon 62. Other basic option fields can be defined and included by the promotions management software 14. The promotion template 50 is stored in a folder in the promotion management software on a computer hard drive of the client or the database for future access by that user or other users, as may be determined by the applicable security policy 56. The process 40a allows the user to produce and store one or more promotion templates 50. From the promotion template 50, other promotion templates can be produced, as well as promotion instances or promotion versions, as discussed below. The security policy 56 determines which users have permission to access the promotion template 50 and what functions the users are allowed to perform with the template (e.g., edit or retire). Security policies govern which users can access which objects (e.g., promotion templates or promotion instances) and the functions those users can perform on those objects (security policies may be applied per object instance or inherited from a parent object or folder) to control access and/or enforce process. The description 58 and suggested uses 60 fields provide further information about the template to aid the user using these templates to produce promotion instances. The icon 62 can be used to help visually identify the promotion template 50. The basic options 52 of the promotion template also can include a promotion code format 64 and a treatment code format 66. The promotion code format 64 and treatment code format 66 are user-specified and each is generated by a code generator customized for a particular system/implementation. The custom code generator includes a custom promotion code generator 68, e.g., "generator1.exe," and a custom treatment code generator 70, e.g., "generator2.exe." When the user produces a promotion instance, the promotion code generator 68 generates a unique promotion code to identify the promotion instance, e.g., "ABC-123" based on the promotion code format 64, e.g., "ABC-nnn."

Referring now to FIG. 6, standard attributes 80 and custom attributes 82 are assigned to the promotion template. A set of standard attributes 80 are included with the promotion management software. These standard attributes 80 are attributes that would be commonly used when producing a promotion template. Examples of commonly used basic options included in the promotion template include a contact channel, a promotion expiration date, and relevant products (used for inferred responses for a promotion). Relevant products can be a list or a query executed against a product table using any attributes within the product table to return a list of unique ProductIDs which can be used to evaluate if the recipient (or member of a hold-out control group) performed the desired action (e.g., purchasing product X).

Custom attributes 82 are user-defined options which can be added to the promotion template. Custom attributes can be of various data types, including string, numeric, date/time, or currency. In addition, each custom attribute can allow free-form entry by the user, provide a fixed drop-down list of values from which the user can select, or provide a drop-down list of values and allow the user to enter additional value(s) which then are subsequently available to all users (modifiable drop-down list). The drop-down list initial values can be

US 8,315,904 B2

5

entered by the administrative user when producing the custom attribute or dynamically queried from a database table or flat file.

In addition to the above, a special custom attribute field can contain multiple values. The multiple values may be manually entered (e.g., multiple string values separated by commas), selected from a fixed list, selected and/or added in a modifiable list, or specified as a query against one or more fields/tables in a database or a flat file. The “related products” field is an example of this type of field where the user specifies a query, typically against a “product table” returning a list of product identifiers (e.g., all stock keeping units where manufacturer=“Fisher-Price™” for products related to the promotion instance. Another example is a multi-valued response channel field, where the consumer must use one of the listed channel(s) when purchasing to qualify as a responder (e.g., purchase is web-based or faxed in). These multiple value fields become important in response tracking attribution (described later).

Standard attributes **80** and custom attributes **82** can either be static (fixed) attributes **84**, static hidden **85**, or parameterized attributes **86**. Static attributes are assigned fixed values when a promotion instance is produced. Parameterized attributes are assigned default values when a promotion instance is produced, but when the promotion instance is assigned to a group in a marketing campaign, its value can be changed. Multiple promotion versions can exist if the promotion instance was produced from a promotion template including at least one parameterized attribute **86**. Each mutation of parameterized attributes for a promotion instance produces a separate promotion version. The promotion attributes can be hashed to determine when a unique mutation exists requiring the production of a new promotion version, where unique instances are stored across campaigns thus reduces storage and computation overhead associated with a normalized history. It is possible to produce multiple promotion versions with the same set of parameterized attributes to optimize run time execution, since promotion versions can still be grouped by their promotion attributes during reporting. A promotion instance produced from a promotion template that does not include any parameterized attributes **86** can have at most one promotion version, that is produced when a campaign is executed, assigning that promotion instance to one or more groups. Control over which attributes are static or parameterized provides a marketing organization control over when new promotion instances are produced versus when promotion versions can be used by controlling which attributes are parameterized. This control provides flexibility in determining how promotions are measured and analyzed for their effectiveness.

Static attributes **84** are constant for the promotion versions produced from the same promotion instance. For example, if a “Fixed-rate home mortgage” promotion template is produced with a static attribute of “closing points,” promotion instance A could be produced with two closing points and promotion instance B could be produced with zero closing points. All promotion versions A1, A2, A3 . . . produced from promotion instance A (by varying other parameterized attributes associated with promotion instance A) would have two closing points. All promotion versions B1, B2, B3 . . . produced from promotion instance B would have zero closing points. The promotion template can specify a default value for the static promotion attribute that can be overridden by the user when the promotion template is used to produce a promotion instance.

Parameterized attributes **86**, on the other hand, may vary in value, producing different promotion versions for each com-

6

bination of values. Again, the promotion template can specify a default value for each parameterized attribute that can be overridden by the user when the promotion template is used to produce a promotion instance. The values specified in the promotion instance for the parameterized attributes **86**, however, are default values that can be changed by the user when the promotion instance is used (i.e., assigned to a group in a marketing campaign).

The static attributes **84** and parameterized attributes **86** can be ordered to control the order in which they are displayed to the user. The standard attributes **80** and custom attributes **82** can be hidden such that when the user produces a promotion instance or promotion version from the promotion template, the hidden attributes **85** of the promotion instance or promotion version cannot be seen by the user or are read-only (displays the attribute name/value, but does not allow user editing). The user can specify default values **88** for the static attributes **84**, hidden attributes **85** and parameterized attributes **86** of the promotion template.

As shown in FIG. 6, examples of default values **88** include “a promotion valid from date” **90**, which can be user-defined or based on the execution of a campaign (that may be implemented as a “flowchart”) separate from the promotion management software **14**. The default values **88** can also be a “promotion expiration date” **92** that is either user-defined absolute dates, or set to expire after a certain period from “the promotion valid from date” **90**, which could be the campaign execution date (e.g., process box run date within the flowchart executing the campaign). Default values **88** can also be promotion value **94** corresponding to a monetary amount, or click-through uniform resource locators (URLs) **96** corresponding to web-based promotions. Any standard attributes **80** or custom attributes **82** that are used in a promotion template as static **84** or parameterized attributes **86** may have default values **88** (hidden attributes **85** are required to have default values **88**). Thus, when the user uses the promotion template to produce a promotion instance, the process **40a** populates the static attributes **84** or parameterized attributes **86** with the default values **88**.

Referring to FIG. 7, process **40a** producing a promotion template includes a user filling in basic options **100**, and specifying the promotion code format **102**. The user specifies a custom promotion code generator **104**. Thereafter, as discussed in detail below, the user specifies a treatment code format **106**. The treatment code format is used to uniquely identify the treatment instance, which is used in contact history and response tracking. The user specifies a custom treatment code generator **108**, which is used in a similar manner as the promotion code generator **104**. The user chooses which standard attributes **80** and custom attributes **82** to include in the promotion template and specifies which promotion attributes are static, hidden, or parameterized **110** and specifies the default values **112**.

Referring now to FIG. 8, the promotion template **50** is used to produce promotion instances **120**, **122** and promotion versions **124**, **126**, **128**. The user modifies the default values of the static attributes in the promotion template **50** to produce promotion instances **120**, **122**. When the user assigns a promotion instance **120** to a group of contact entities, the user specifies different sets of values for the parameterized attributes in the promotion instance **120** to produce promotion versions **124**, **126**, **128**. The group of contact entities can either be target groups of contacts receiving promotions, or control groups of contacts to test promotion effectiveness. Assigning promotions to groups is generally performed within a contact-style process box (not shown) of the software. A contact-style process box facilitates assigning pro-

US 8,315,904 B2

7

motions to groups over specific contact channels in an execution of a marketing campaign. Contact channels include direct mail, contact centers, branches/stores, or email. For the user's convenience, multiple promotions can be assigned to multiple groups within a single contact-style process box, designating a contact channel. Contact-style process boxes are discussed in further detail below with regard to contact history.

The user produces promotion instances **120**, **122** when the user changes the value of one or more static or parameterized attributes. These are treated as separate promotion instances in the promotion management software. Promotion instances **120**, **122** have different promotion codes to distinguish one promotion instance from another, even though the promotion instances may have the same name and/or attribute values.

The user produces promotion versions **124**, **126**, **128** when the user changes the value of one or more parameterized attributes, as defined in the promotion template **50**. The promotion versions **124**, **126**, **128** have the same static attribute values and the same promotion code. The promotion management system treats promotion versions **124**, **126**, **128** as mutations of a common promotion instance, for example, promotion instance **120**. The feature of parameterized attributes allows a user to conveniently produce a plurality of promotion versions having the same values for static attributes (including basic options).

Referring to FIG. 9, to illustrate the above features, promotion template **130** is shown having basic options of a promotion template name **132** of value "College Card," and a promotion code format **134** of value "ABC-mnn." The promotion template **130** also has standard attributes, which are a static attributes, namely a description **136** of value "College Logo Card" and contact channel **138** of value "Email." There is also a standard attribute which is a parameterized attribute, namely, expiration date **140** of default value "6/30/05." Promotion template **130** also has custom attributes which are parameterized attributes, namely, college **142** of default value "Harvard," and promo rate **144** of default value "5.25%."

The user produces promotion instance **146** from the promotion template **130** using all of the default values. Based on the promotion code format **134**, the promotion code generator generates the unique promotion code "ABC-123." The user assigns promotion instance **146** to a group of Harvard graduates, producing a promotion version **148** with the same values as the promotion instance **146**. The user assigns promotion instance **146** to a group of Princeton graduates, but changes the default values of parameterized attributes expiration date **140** and college **142** to produce promotion version **154**. Specifically, when the user produces promotion version **154**, expiration date **140** of default value "6/30/05" is changed to value "9/30/05," and college **142** of default value "Harvard" is changed to value "Princeton." New values "9/30/05" and "Princeton" are assigned to expiration date **150** and college **152** of promotion version **154**, respectively. The other static attributes of promotion version **154** remain the same as the corresponding attributes in promotion instance **146**. Both promotion versions **148**, **154** have the same promotion code of value "ABC-123." Thus, promotion versions **148**, **154** are mutations of a promotion instance **146**.

The user can re-use the promotion template **130** to produce a promotion instance **156** with different attribute values. Here, the user has changed the value of static attribute "contact channel" **138** from "Email" to a new value "Mail." Values of all other attributes remain unchanged. When the user produces the promotion instance **156**, the promotion management software treats promotion instance **156** as a separate

8

promotion instance from the promotion instance **146**. There is a unique promotion code **158** of value "ABC-124" for promotion instance **156**.

Promotion templates can be used to produce new promotion instances when the promotion template is "active." A promotion template, once it has been used to produce one or more promotion instances, can be "retired" (but not deleted). Retired promotion templates can no longer be used to produce additional promotion instances, but are still available for reporting and system table integrity.

Referring to FIG. 10, while the user can manually populate the values of the parameterized attributes with a constant value, as discussed above, the values of the parameterized attributes **160** can also be populated from a database or list, e.g., database **24** (FIG. 1) or by computing values based on data from a database (i.e., calculated fields). Populating the values from data may be desirable to support a large number of promotions versions to be assigned to target groups. "Data-driven" parameterized attribute fields **162** are, e.g., "FavoriteTrips.From" and "FavoriteTrips.To" and correspond with the "From" and "To" columns of the "driving table" (Favorite Trips) in the database. The data in the driving table is matched to the target entities by a code (e.g., Customer ID) and each row of the driving table corresponds to one promotion version given to the target recipient. Thus, by mapping the data to the data-driven parameterized attribute fields, the promotion management software **14** can automatically produce, at run time of a campaign flowchart, the promotion versions needed, without the user manually producing promotion versions, as discussed above.

Since any target entity can have a variable number of rows associated with it in the driving table, data-driven parameterization provides truly data-driven personalization of promotions. For example, customer #3254678 is offered three different flight promotions (Boston to New York, Boston to Orlando, and New York to London) while customer #8452353 is offered four flight promotions (Atlanta to Chicago, Atlanta to Los Angeles, Atlanta to San Diego, and Atlanta to Boston). This solution provides tremendous scalability in one-to-one promotion personalization, while retaining full granularity of promotion information for tracking purposes (to be discussed below).

Custom validation of standard and custom attributes of promotion instances and promotion versions can be implemented to ensure that the combination of values chosen for a particular promotion instance or version meets the requirements of the company. These validation checks could verify numeric/date/currency ranges, compare attribute values, enforce specific dependencies between multiple fields, provide security (which user is allowed to make certain types of changes).

Referring to FIGS. 11 and 12, a group management process **40b** that manages the assignment of promotion instances to groups is shown. Groups, and in particular target groups, are collections of contact entities to be targeted in the marketing campaign. Group management process **40b** can be implemented as either a top-down campaign definition or a bottom-up campaign definition. Top-down campaign definition refers to defining a marketing campaign at a conceptual level (a description of the marketing campaign the user intends to implement). A top-down campaign definition can support, for example, a marketing planner who has access to volumes of raw data and is better suited to plan the marketing campaign at a conceptual level. This approach also allows promotions to be produced before or after the targeting criteria have been established. A bottom-up campaign definition refers to defining a marketing campaign at an implementation level and

US 8,315,904 B2

9

provides a visual representation of the treatments (promotions assigned to target groups) in a marketing campaign. Thus, bottom-up campaign definition can support, for example, a technical marketer to augment an existing marketing campaign with additional promotions derived from day-to-day interaction with contact entities. Combined, the two-level planning and review structure of both the top-down and bottom-up campaign definitions can serve as quality assurance for the marketing campaigns as a whole.

Referring to FIG. 11 in a top-down campaign definition, the user produces a structure to manage records 170. One such structure is a target group spreadsheet. Other structures could be used, such as a table. The structure, as a target group spreadsheet includes rows representing the target groups in the marketing campaign. The rows log what promotion instances have been assigned to which target groups of contact entities. When the user desires to add to the marketing campaign, the user generates a row 172 in the target group spreadsheet, assigns one or more promotion instances to the target group. The user can assign the promotion instance and leave the specification of the parameterized attribute values to the user implementing the marketing campaign, or he can specify one or more of the parameterized attributes to be used with each promotion instance.

The user also specifies group attributes of the target group in the target group spreadsheet. The group attributes are customizable and represent information about the target group that may be output as part of the contact list generated by the marketing campaign, the selection criteria or definition of the target group, or used for data analysis/reporting. Group attributes can be a group name/category, a unique group code to identify the target group, the minimum and/or maximum group size, the group source identifier, the age, income, recency, frequency, or monetary values defining the target group. When the custom group attributes represent selection criteria, these fields and their values can be made available in the campaign/flowchart to implement the selection logic without duplicate data entry. These group attributes are entered into the row.

The user determines whether a control group should be assigned to the target group 174. As explained above, the hold-out control group represents marketing entities that would be statistically representative (e.g., similar in demographic or financial bracket) to the marketing entities of the target group. The user uses the control group to test the effectiveness of the marketing campaign (i.e., measure the incremental benefit of sending the promotion, also known as "lift" over the control group). If the user determines that a control group is desired, then the user assigns a control group to the target group 176. The control group is measured against every promotion instance assigned to its target group. A control group can be used multiple times for different target groups if desired. Next, the user determines if group-level custom attributes are desired 178. The order of determining whether a control group is desired 174 and determining if group attributes are desired 178 can be reversed if desired in the group management process 40b.

Otherwise, if a control group is not desired 174, then the user continues to determine if group-level custom attributes are desired 178. If the user determines that group-level custom attributes are desired 178, then the user assigns one or more group-level custom attributes to the target group 180 and enters values for the custom attributes in the target group spreadsheet. Selected group-level custom attributes are available for every row in the target group spreadsheet. Then the user determines if there are more promotion instances to be assigned to target groups 182. Otherwise, if a group-level

10

custom attributes are not desired, then the user determines if there are more promotion instances to be assigned to target groups 182.

Group-level custom attributes are, in addition to the promotion attributes described above. Attributes are assigned to the target group independently from when the promotion instance is assigned to the group (they can be assigned/filled in at any time after the target group row has been produced). Group-level custom attributes can be assigned to both target and control groups. The group-level custom attributes represent information specific to the group. For example, group-level custom attributes can be the target group size representing the minimum or maximum number of contact entities in the group. The group-level custom attributes can also be a group description or a designation that the group is of control group type or target group type. The user can also specify the audience type of the contact entities of the group, whether the type be a cross-sell, upsell, defection, loyal, or retention. The group-level custom attributes can also be an audience affinity, designating the group as contact entities within a company division, or an audience source, designating where the source of the contact entities. The audience source can be derived from a rental list, owner list, prospect list, or a lead list. The group-level custom attributes can also designate a lead source, the lead source being television, road-show, radio, or direct mail.

Still referring to FIG. 11, if the user determines that there are no more promotion instances to be assigned to target groups 182, then the process produces one or more campaign flowcharts 184 (i.e., the implementation of the marketing campaign organized into different groups of execution logic for convenience if desired) using the target group definitions in the target group spreadsheet. In the illustrated example of FIG. 13, the campaign flowchart 220 allows a low-level user, such as a technical marketer to implement the marketing campaign based on the definition provided by the marketing planner in the target group spreadsheet 222 (in some cases, the marketing planner and the technical marketer may be the same individual). The flowchart instances have access to the group-level custom attributes of the rows of the target group spreadsheet.

As shown in FIG. 11A, the process of producing a flowchart from the target group spreadsheet 184 can be broken down into the following steps. When the technical marketer produces a target group in the flowchart, he can then associate it with the appropriate top-down entry in the target group spreadsheet. This association causes the group in the flowchart to automatically inherit all of the attributes of the target group defined in the target group spreadsheet, including the assigned promotion instances, control group, and group-level custom attributes. The technical marketer initiates the target group association process 184a. This displays a list of the target groups entries from the target group spreadsheet (this can be filtered to show only the entries that have not yet been associated with another group of a flowchart). The technical marketer then chooses a target group entry to associate with the current flowchart group 184b. The group automatically inherits the target group spreadsheet entry's group name and code. In addition, when the technical marketer configures the treatment of that flowchart group 184c, the promotion instances and control group, and group-level custom attributes from the target group spreadsheet entry are automatically applied.

The user saves the target group spreadsheet and/or the campaign flowchart 186. Thereafter, the user can edit the information in the rows of the target group spreadsheet and in the flowchart instances of the campaign flowchart. If there are

US 8,315,904 B2

11

changes to the target group spreadsheet **188**, then the process mirrors the changes in the campaign flowchart **190**. Similarly, if there are changes to the campaign flowchart **192**, then the process mirrors the changes in the target group spreadsheet **194**. The user saves the changes to the target group spreadsheet and/or campaign flowchart **186**. In some implementations, the user can make multiple changes to either the target group spreadsheet or the campaign flowchart before saving. However, the process will mirror the changes after a save.

Referring to FIG. 12, in a bottom-up campaign definition, the user produces an overall representation of a campaign **196**, which could be visual, such as a nodal diagram, graph or as will now be discussed, a campaign flowchart. The campaign comprises flowchart instances **198**, the flowchart instances being visual representations of the execution logic used to produce the intermediate groups and final target groups (e.g., to handle selection, suppressions/exclusions, segmentation, sampling, audience-level switching or grouping contacts into households). The flowchart instance also includes what promotion instances have been assigned to which target groups of contact entities, as well as control group assignments. When the user wishes to add to the marketing campaign, the user generates a flowchart instance in the campaign, assigns one or more promotion instances to one or more target groups, and/or specifies group attributes of the target group(s). These group attributes can be entered into the flowchart instance (or in the target group spreadsheet).

The user determines whether a control group should be assigned to each target group **200**. If the user determines **200** that a control group is desired, then the user assigns a control group to the target group **202**, and determines if group-level custom attributes are desired **204**. Otherwise, if a control group is not desired, the user determines if group-level custom attributes are desired **204**. If the user determines that a group-level custom attributes are desired **204**, then the user assigns one or more group-level custom attributes to the target group **206** and enters values for the custom attributes in the flowchart. Then the user determines if there are more promotion instances to be assigned to target groups **208**. Otherwise, if group-level custom attributes are not desired, the user determines if there are more promotion instances to be assigned to target groups **208**.

If the user determines that there are no more promotion instances to be assigned to target groups **208**, then the process produces a target group spreadsheet from the campaign flowchart **210**. The target group spreadsheet comprises rows with group attributes which are populated from the group attributes of the flowchart instances of the campaign, i.e., the bottom-up entries, in addition to any top-down produced entries.

The user saves target group spreadsheet and/or the campaign flowchart **212**. Thereafter, the user can edit the information in the rows of the target group spreadsheet and in the flowchart instances of the campaign flowchart. If there are changes to the target group spreadsheet **214**, then the process mirrors the changes in the campaign flowchart **216**. Similarly, if there are changes to the campaign flowchart **218**, then the process mirrors the changes in the target group spreadsheet **220**. The user saves the changes to the target group spreadsheet and/or campaign flowchart **212**. In some implementations, the user can make multiple changes to either the target group spreadsheet or the campaign flowchart before saving. However, the process will mirror the changes after the save.

The top-down and bottom-up campaign definition processes are described separately for ease of comprehension, but these two processes can be arbitrarily intermixed. That is, some target groups can be defined top-down and associated

12

with flowchart groups, some can be defined top-down and not used by any flowchart group, and some can be defined bottom-up. All three of these states are visually differentiated in the target group spreadsheet to facilitate understanding and verification of the entire campaign development process.

Having defined the marketing campaign by assigning promotion instances to target groups, the marketing campaign can be further customized and personalized to target the specific needs or desires of the contact entities in the target groups. Thus, promotion instances can produce promotion versions, which are mutations of the promotion instances depending on variations of parameterized attributes of the promotion instances.

Referring to FIG. 14, the execution of the marketing campaign process **40c** (in production as opposed to a “test run”) produces a treatment instance **240**, which is the intersection of a target group **242** (marketing entities being promoted to or treated, which may be a specified list of audience IDs or a group with unspecified details (e.g., mass marketing where the count of target entities may be known, but the specific identities are not)), a promotion version **244** (of promotion instance **246** given to the group), and a run time **248** (the specific date/time the marketing campaign flowchart is run). When the user runs the marketing campaign process **40c** and assigns the promotion instance **246** to the target group **242**, the user specifies the values of the parameterized attributes, as discussed earlier, producing promotion version **244**.

Once the campaign is run and the treatment instance is produced, treatment code generator **68** (FIG. 5) generates a unique treatment code to identify the treatment instance **206**, e.g., “ABCDEFGH” for the treatment instance, based on the treatment code format **64**, e.g., “aaaaaaaa”. The user can clear (delete) contact history by the run instance by viewing and selecting from a list of run date-times (this is useful if a marketing campaign was run in production in error or it was later decided not to send out the promotions from the campaign).

Additionally, once the user produces a promotion instance, the user can retire the promotion instance. Retiring prevents further use of the promotion instance, so that the user cannot assign the retired promotion instance to any groups and any group already assigned the retired promotion instance will not be marketed (e.g., the flowchart will generate an error when run). The promotion instance no longer appears in the promotion tree hierarchy. However, it is still available for reporting/analysis purposes.

Referring now to FIGS. 15 and 15A, once a marketing campaign is run **250**, contact history process **40d** of the promotion management software produces **250a** (FIG. 15A) tables of marketing campaign data to maintain a contact history of the campaign. Contact history process **40d** provides a historical record of every communication made to a marketing entity providing detailed information on exactly what was given to whom, when, and how (in order to track responses). It also includes tracking of every contact entity held out from a promotion in a control group (required for inferred response tracking). Contact history may be written at the time that contact lists are generated (e.g., when the flowchart runs) or they may be deferred and written later (e.g., after an approval process or confirmation that the delivery was completed).

Contact history can provide an accurate way to count the number of “interruptions” to a contact entity in view of the number of promotion instances assigned to the contact entity (entries where an individual is part of a hold-out control group are automatically eliminated from this count based on the entry’s contact status). Accurate management of interruptions is critical for marketers to prevent customer fatigue

US 8,315,904 B2

13

(negative behavior, such as opting out from marketing communications or a decrease in response behavior) when the frequency of communications is too high. If the user wishes to limit the number of contacts per entity over a given period of time, logging contact history according to packages can serve to help limit contact fatigue by accurately tracking the number of interruptions. In addition, since contact history can grow extremely large for an organization with many customers or prospects, the ability to compress this data without loss of information is desirable. The combination of base contact history and detailed contact history provides this compression benefit whenever possible (i.e., all members of the same target group receive the same treatments), without losing granular data when members of the same target group receive different promotions.

Interruptions may refer to the number of actual communications, or “packages,” made to contact entities over a given contact channel and can contain multiple promotion instances. A mechanism is provided in the campaign management solution to define which contacts are considered part of the same package. For example, a package might be defined as all of the treatments generated within a single contact-style process box corresponding to a marketing campaign.

For example, suppose a contact entity receives multiple promotions in a marketing campaign. The multiple promotions can be sent over the same contact channel, e.g., by direct mail, or over multiple contact channels, e.g., a combination of direct mail and electronic mail. In order to track the number of interruptions made to the entity, the process 40d treats the multiple promotions assigned through the same contact-style process box as a single interruption, or “package,” as opposed to treating each of the multiple promotions assigned as a separate interruption. Thus, in the case where the user promotes fifty different promotions to a target group in a contact-style process box, i.e., in the same envelope via direct mail, each of the fifty promotions are considered part of the same package, and constitute a single interruption to each marketing entity. In the case where the user promotes thirty promotions over direct mail, and twenty by electronic mail, then by this definition, there are two packages since the user is interrupted twice.

Contact history 252 includes a promotion version history table 254, a treatment history table 256, and a base contact history table 258. Each table comprises an aggregate of marketing campaigns executed over time (of both targeted groups and control groups). There is a separate contact history table for each audience level (though these could populate the same underlying physical database table). Audience levels represent different types of marketing entities that can be contacted, including the household level, the customer level, and the account level. In this hierarchy, households may contain one or more individuals (customers) and customers may hold one or more accounts (many-to-many relationship).

The promotion version history table 254 is a record of all promotion attribute values for each promotion version 266 for all marketing campaigns (including the campaign 250) when they are executed. Each promotion version 266 is assigned a unique promotion version number 264. The promotion version history table 254 captures the values for both parameterized and static attributes as assigned. A variation of the promotion version history table can capture the parameterized attributes, which can save tremendous storage space as static attributes typically do not change. Static attribute values can then be retrieved directly from the promotion instance definition (furthermore, the application could prevent the changing of any static promotion attribute values after a promotion instance has been used in any campaign).

14

The promotion version history table 254 can be linked 250b (FIG. 15A) via pointer 268 with the treatment history table 256 through the promotion version numbers 270. The treatment history table 256 records 250c (FIG. 15A) treatment instances 270 produced when the marketing campaign is executed 250, as discussed above. For each treatment instance 270, the treatment history table 256 records 250d (FIG. 15A) the groupID 272 identifying the target group of the promotion version (as enumerated by the promotion version number 270). The run date and time of the marketing campaign 274 also recorded 250e (FIG. 15A). Each treatment instance 270 is assigned 250f (FIG. 15A) a globally unique treatment code 276 for exact identification of the specific marketing instance for tracking responses. Rows in the treatment table also could have custom attributes for tracking additional treatment details (e.g., estimated size of the target group for brand marketing communications).

For every target group that is assigned a control group, there is corresponding control treatment instance produced 20 for every treatment instance (and subsequently linked to the treatment instance). For example, if target group G is assigned two offers A and B and is assigned control group C, the treatment history table 256 records two treatment instances G-A and G-B (for group G receiving offer A and B) 25 and two control treatment instances C-A and C-B (for group C receiving offer A and B) where C-A is linked to G-A and C-B is linked to G-B.

The treatment history table 256 can be linked 250g (FIG. 15A) 268 with the base contact history table 258 through the groupID 272 and the package ID. When the marketing campaign is executed, the base contact history table 258 records 250h (FIG. 15A) a row 278 to track which contact entities, identified by customer level custID 280, were assigned to what groups, identified by groupID 272, along with the contact date and time 282. The contact date and time 282 is initially set to the date and time the marketing campaign is executed, but the date and time 282 can be updated to reflect subsequent updates to the record (e.g., updating the contact status from “Attempted delivery” to “Undeliverable”) made 40 with the contact entity, as described in further detail below.

In a similar manner, the base contact history table 258 also tracks the entities assigned to control groups. The storage of group membership information in the base contact history table 258 is highly compact when all members of the same group receive the same promotion versions. When this is not the case, the exact contact information is stored in detailed contact history (discussed below).

As discussed earlier, a “package,” designated in the base contact history table 258 as packageID 284, represents a single “interruption” of the contact entity. This can be represented as the multiple promotion versions assigned in a contact-style process box over a single contact channel a marketing campaign, or it can be a value explicitly assigned by the user to group together multiple promotion instances/communications. For example, as shown in the treatment history table 256, promotion version number 113 is assigned to five groups, C201, C202, C203, C204, and C205 when a marketing campaign is executed on Jun. 1, 2005. Although the customers of the five groups are respectively labeled with 55 groupIDs of C201, C202, C203, C204, and C205, as shown in the base contact history table 258, the packageIDs for those entries are the same, namely, “78904.” If a marketing entity is a member of more than one of these groups, the promotion versions from the same contact-style process box assigned to 60 their respective entities are considered one interruption.

The rows 278 of the base contact history table 258 also include a contact status (not shown). The contact status pro-

US 8,315,904 B2

15

vides flexibility in tracking the status of a contact entity that has been promoted. Each contact status defines whether or not it counts as a contact/interruption to the recipient. For example, a promotion version “served” to a call center representative’s screen that has not yet been presented to the caller does not count as a contact. Once the promotion version has been “read” to the caller, it may have a change of contact status to “Presented” which does count as a contact. In some implementations, contact status is conducted at the package level. In such cases, the user updates contact status for all promotions within the package.

The user can customize the status entries associated with the contact status. For example, as shown in FIG. 15B, contact status tree 300 can track the status of a promoted contact entity. The contact status tree can be structured so that the user understands the hierarchical meaning of a particular contact status (i.e., there is only one path to any given status). For example, a row 278 could have a status of “served” when the promotion version is assigned to the target group and the marketing campaign is executed. If the user promoted the promotion over email and, for instance, requested a read receipt, upon return of the read receipt, then the contact status could be updated to “confirmed contact.” The update date and time 282 would be updated accordingly (the contact date and time remains the same). If, by chance, a second user reviews the base contact history table and views the “confirmed contact,” then, by virtue of the hierarchical structure of the contact history tree, the second user understands that “contact initiated,” “served” and “proposed” contact statuses were necessarily met. By the same token, the second user knows that the contact could not have been “dropped.” It is possible to have customized contact statuses where the system has no knowledge of any hierarchy and contact history simply tracks the current contact status value. For each contact status value, the system can configure whether the contact status actually “counts as a contact” or not (e.g., “proposed” and “undelivered” would not count as contacts, while “contact initiated” and “contact confirmed” would).

It is also possible to append additional rows to contact history to effectively update the contact status, allowing the recording of the specific date and time of each contact status change, though this can cause contact history to grow significantly. A special contact status is used for records written to contact history for control groups (and thus can be easily filtered out when computing “interruptions”).

In some implementations, contact history includes additional tables (not shown) such as a group history table, a campaign history table, a strategic segment history table, and a detailed contact history table. The group history table logs changes to the target group spreadsheet entries (e.g., changes to group custom attribute values over time). The campaign history tracks changing values of campaign attributes over time (e.g., custom campaign attributes). The strategic segment history stores information about how the company divides the contact entities into groups (i.e., records the membership of marketing entities within strategic segments such as high, medium, or low-value customer segments) over time. Thus, the process allows the system to accurately track history and information for a variety of aspects related to a marketing campaign so that at any given time, the system can reconstruct the state of any marketing object within the system when a communication was made to a contact entity.

The detailed contact history table is similar to the base contact history tables. There is (conceptually) one table per audience level. Detailed contact history tables are used in special cases to track promotions given to contacts on a data-driven parameterized promotion attribute basis (i.e., when

16

promotions are personalized to the contact entity within a target group). This includes the tracking of detailed control group information, where each recipient is held out as a control for the specific promotion versions he would have received if he had not been part of the control group. Detailed contact history tables also track promotions that have been rule optimized to deter contact fatigue. Both of these cases track detailed history in the case where all members of a target group do not receive exactly the same promotion versions. Members of the same target group may receive different numbers of promotions and different promotions, personalized based on the specific contact entity’s data. Detailed contact history tables also track promotions assigned to contacts on a real-time basis, where tracking the number of promotion presentations per contact entity is important.

Referring now to FIGS. 16 and 17, a promotion list process 40e to produce promotion lists is shown. A promotion list is a collection of one or more promotion instances and can be used where ever individual promotion instances are used (e.g., for promotion assignment to target groups). The promotion list includes promotion list attributes 350. Promotion list attributes 350 include a unique promotion list name 352 and a promotion list description 354. The promotion list can be a static promotion list or a smart promotion list, both of which are described below. The promotion list is stored on an electronic medium, e.g., the server system hard drive or application database, and can be viewed in a folder tree view 356 (this view can be intermixed with promotion instances). Promotion lists 358 and promotion instances are assigned an icon 360. Thus, when promotion lists are viewed in the folder tree view of FIG. 17, the icon 360 can be used to identify the type of promotion list, e.g., whether the promotion list is a static promotion list 362, a smart promotion list 364, or a single promotion instance 366.

Referring now to FIG. 18, a static promotion list is produced by using a searchable and browsable view (e.g., a folder tree view) 356 of the promotion instances 368 stored in folders 370 on the hard drive of the server 12 or application database 22 (see FIG. 1). The searchable and browsable view allows the user to search, view in a dynamic nested tree, drill into the folder pages, and browse thumbnail views and saved searches. Search for promotion instances can be based on any promotion attributes as well as “inherent attributes” such as the creation or update date/time, owner, folder, promotion template, and security policy. Based browsing or the results of the search, the user selects one or more promotion instances to add to the static promotion list. Depicted are promotion instances within the “Recently Used Offers” folder. The user selects promotion instances “No Money Down E-mail,” “15% off Holiday Mailer,” and “Free Shipping Offer” from the folder. These promotion instances can be ordered up and down by the user to select their display order/rank. The promotion instances of such a static promotion list will not change until the user repeats this process to add or delete promotion instances to the static promotion list.

Referring now to FIG. 19, a smart/dynamic promotion list can be produced by querying 372 promotion instances to find promotion instances with matching search criterias 374. The search criteria can be based on promotion attributes, (e.g., Contact Channel, Expiration Date, Cost per offer), other inherent attributes (e.g., owner, creation date, update date, folder, promotion template) and operators, e.g., ($>$, $<$, \geq , \leq , $=$, \neq , begins with, contains, not contains, not begins with). Multiple search criteria can be connected via operators (e.g., AND, OR).

The user constructs a query 372 to send to the search process that searches the hard drive of the server 12 or appli-

US 8,315,904 B2

17

cation database 22 for stored promotion instances to retrieve. The retrieved stored promotion instances will have an owner (CreateBy field) with value of “rkennedy,” “AND” channel with value of “Email,” as shown in FIG. 19. The “AND” is an example of using a search term that is a Boolean operator to cause the process to include both “rkennedy” and “Email” as search criteria 374. The user can limit the search to specified folders and optionally their subfolders 376. The results can be sorted 378 in ascending or descending order according to one or more specified fields. The results can also be limited 380 to the first N number of promotion instances as ranked by the sorting criteria 378.

The sorting criteria 378 can be based on one or more attributes of the promotion or a calculated field based on attributes of the promotion and multiple attributes/calculated fields may be used to sort based on ties (e.g., select first in descending order based on promotion instance creation date and then for any on the same date, then subsequently sort in ascending order by promotion cost).

The sort 378 and limit 380 features can be useful in limiting the results to a maximum or minimum case. For example, suppose the user wishes to query all promotion instances for gold credit cards, but only wants the promotion instance for the gold credit card with the lowest interest rate. Thus, if the user sorts 378 results of promotion instances for gold credit cards in ascending order by interest rate, and limits 380 the results to the first hit, then the user can find, the promotion instance for a gold credit card with the lowest interest rate. Another example might be to return the top three promotion instances in a folder ranked in descending order by offer score. The marketing communications to a particular group targeted by a monthly marketing campaign can then be managed (e.g., maintained, updated, rescheduled) without having to actually edit the campaign (e.g., by simply moving offers in and out of this folder). This is a powerful concept in producing periodically recurring marketing campaigns that can run in a lights-out (i.e., without manual intervention) manner, dramatically speeding up an organization’s ability to scale and maintain a large number of campaigns.

If the user produces a smart promotion list, the query 372 will dynamically return matching promotion instances at the time the smart promotion list is used (e.g., when a campaign is executed and promotions are assigned to a group via a smart promotion list). Thus, without further user intervention, the process updates the promotion instances of a smart promotion list according to the query 372. For example, as shown in FIG. 20, the promotion instances 380, 382 for a gold credit card with the lowest interest rate 384, 386 may vary day-to-day according to fluctuating interest rates. A smart promotion list 388 can be updated without further user intervention to reflect such fluctuations.

Another key example for smart promotion lists is in building optimization rules based on promotion meta data (e.g., limiting the number of loan offers an individual may receive to three a quarter). The optimization rules regulate the use of limited marketing resources or limited audience tolerance for marketing interruptions, to optimize, for example, performance or profitability. As new promotion instances are produced that are credit vehicles, they would automatically be picked up by the smart promotion list based on specified query conditions.

Once a marketing campaign has been executed, the user can then run a response tracking process to track and assess the performance of the marketing campaign and the promotions. The response tracking process 40/ tracks duplicate and unique responses. The response tracking process also tracks responses from customers that were originally targeted by the

18

marketing campaign, or alternatively, responses from customers as a result of viral marketing, i.e., customers who come to know of promotions as a result of word of mouth or pass-alongs and provide a system-generated code (e.g., campaign, promotion, group, or treatment code) with their response. The response tracking process also tracks responses before or (for some configurable time period) after the expiration date of the promotion. The configurable time period may be a single value for the entire system or it may be different values depending on the promotion type, the promotion channel, etc. (e.g., an organization may want to track “late” responses for email offers for a period of three days, telemarketing offers for a period of seven days, and direct mail offers for thirty days).

Referring to FIG. 21, the response tracking process 40/ receives events or actions performed by audience members that may represent direct or indirect responses to promotion instances, as reported 400 to the system through various contact channels such as direct mail, contact centers, branches or stores, or eMessaging, e.g., electronic mail. The responses types 402 can vary from inquiries related to the promotion version of a treatment instance, to sign-ups for trial periods for goods or services related to the promotion version of a treatment instance, to actual purchases for goods or services related to the promotion version of a treatment instance.

The information reported 400 can be entered into an action table 404, one per audience level. Each row of the action table (an event) includes multiple fields to be used to log the data associated with each response. For example, action table fields include the audience field(s) associated with contacted entity making the response, and the date and time of the response. The event also identifies the “response type,” system-configurable categories of responses (e.g., click-throughs, purchases, account activation, inquiry, etc.). Complex responses requiring a sequence of actions also can be handled by submitting an appropriate event and response type (e.g., purchase of two Disney™ digital versatile discs (DVDs) after purchasing The Lion King™ to the action table, based on detecting the presence of the appropriate purchase transactions).

Each response type definition indicates whether it counts as a response (i.e., whether it rolls up as a response in performance reporting). For example, while a marketing user may track the number of inquiries for a credit card promotion, the number of applications, the number of subsequent activations, and the number of times the card is used within the first month post-activation, only some of these response types are typically identified as counting as a response for roll-up or financial return on investment calculations (e.g., inquiries may not count, but card activation may count; or click-throughs may not count, but purchases may count).

The action table 404 also includes fields for the possible response codes (e.g., a campaign code, a promotion code, a group code, or a treatment code). These response codes are used to help identify one or more treatment instances that were promoted to the contacted entity. For example, the response code could be printed on the coupon or other promotion item, and would be included in the data associated with the response. Other fields in the action table 404 include fields for promotion attributes of the promotion, and these attributes can be either static, parameterized, or both. For any of these fields, the user enters the data known from the response. It is possible that some information is not known, and the fields for the unknown information are null.

The fields in the action table 404 also include a response type field. The response types may include a click-through, an inquiry, opening an account, a purchase, activating an

US 8,315,904 B2

19

account. Each row of the action table (an event) is associated with one response type. Therefore, multiple responses to the same promotion, for example, an inquiry followed by a purchase, are two events (i.e., two row entries) in the action table corresponding to the two response types. In other words, responses associated with multiple action types, even responses from the same contacted entity on the same promotion, are entered into the action table **404** multiple times. (If a response type hierarchy is defined and there is only one path to any given response type, multiple entries are not required for different response types and the occurrence of response types prior to the currently reported response type can be inferred).

The fields also include a product type field, corresponding to code or other identifier of the goods or services promoted in the marketing campaign, which can be compared to a list of products associated with a promotion instance for inferred responses.

When the response events are entered into the action table **404**, the responses are attributed **406** to one or more treatment instances. As discussed earlier, treatment instances correspond to a promotion version assigned to a group of contacted entities during the execution of a marketing campaign. Attributing responses to one or more treatment instances credits one or more treatment instances for triggering the response. Attributing **406** includes identifying **408** one or more possible treatment instances to receive credit for the response, and distributing **410** the credit between the one or more designated treatment instances according to a set of rules, and recording relevant information about the response **412**, such as whether the respondent was in the originally contacted group and/or whether the response date was on or prior to the promotion expiration date.

Attributing **406** includes identifying **408** one or more possible treatment instances to receive credit for the response includes determining **410** whether the response is one of two basic types of responses: a direct response or an inferred response. For a response to be a direct response, the response information received includes one or more response codes. In other words, a direct response includes a system-generated code that was sent out with the promotion (e.g., either a campaign code, a promotion code, a group code, or a treatment code). These codes are matched to treatment instances through lookups of the treatment history table or other contact history tables. An inferred response does not include response codes and can only be credited if the treatment instance contacted the responder. In both direct and inferred responses, promotion attributes reported with the response are used to further assist in identifying one or more possible treatment instances to credit. Similarly, if a product code is reported with the response, the product code can be matched against a promotion instance's relevant products in a product table, the products being goods or services promoted in the marketing campaign. The promotion attributes and product codes, if reported with the response, are therefore important in identifying treatment instances in the inferred response case.

In some implementations regarding the direct response, if the treatment code is known with the response, then an exact and unique treatment instance can be identified (without any further information). This is a direct exact response. Otherwise, if the treatment code is not reported but other response codes are reported, it is foreseeable that multiple possible treatment instances are identified. This is a direct inexact response. For example, if the campaign code is known, then the range of possible treatment instances can be narrowed to those treatment instances with promotion versions executed in the marketing campaign associated with the campaign

20

code. However, if the campaign contained multiple promotion instances and/or ran multiple times, the exact treatment instance is not known from the campaign code alone. In another example, if the promotion code is known, then the range of possible treatment instances can be narrowed to those treatment instances with promotion versions of promotion instances associated with the promotion code.

In some implementations regarding the inferred response, if no response codes are known with the response, then a unique treatment instance or even multiple possible treatment instances are not immediately identified. Inferred response cases can represent instances of "viral marketing" where the marketing campaign did not target the customer but, for example, through word of mouth, the customer learned of the promotion and took the desired behavior. In such cases, the user relies on other data fields in the action table to infer multiple possible treatment instances. For example, the response would likely contain information regarding the product or custom promotion attributes associated with the promotion. Using the information known from the response, the user can infer that one or more treatment instances were responsible for eliciting the desired behavior and can therefore be credited.

Referring to FIGS. **21A** to **21C**, the response tracking process **40f** retrieves and processes one row (event) at a time from the action table, until all rows have been processed **450** (though described here conceptually as processing a single row at a time, response processing is not limited to a single row at a time as all rows can be processed simultaneously in a batch process). If one or more response codes have been provided **422** (direct response), then any non-NULL codes and promotion attributes are used to match against treatment instances to find the set of possible treatment instances to receive credit for this response event **424**. From the set of possible treatment instances, if any treatment instance contacted the responder **426**, then all treatment instances that did not contact the responder are discarded **428**. If no response codes are provided in the event **422** (inferred response), then the provided promotion attributes are matched against only treatment instances that contacted the responder **430**.

From the set of possible treatment instances, if the response date is between the effective and expiration dates of any treatment instances **432**, then discard any treatment instances in which the response was not within the valid date range **434**. For the remaining treatment instances, record all the appropriate response tracking information using multiple, simultaneous attribution methods.

Attributing responses to one or more treatment instances **436** includes distributing the credit among the one or more remaining treatment instances according to a set of rules. The rules include best match, fractional attribution, multiple attribution and custom attribution. In addition, response attributes are recorded identifying whether the responder was within the originally contacted target group or not, and whether the response date was within the valid promotion effective and expiration dates. Data also is captured to allow the user to count the number of total responders and/or the number of unique responders. Additionally, the marketing user can choose to track other responder state information (e.g., the response channel or the current credit card balance after a purchase stimulation campaign) as additionally tracked fields.

Once the attribution methods and response attributes have been captured, the user can choose to view any combination of attribution method(s)/attribute(s) for any response type(s) in performance reports. For example, one report could show different columns for 1) the number of unique individuals that

US 8,315,904 B2

21

were originally contacted that purchased (purchase is an example of a response type) prior the expiration date, 2) the total number of responses by individuals that were originally contacted that purchased prior the expiration date, and 3) the total number of purchases by individuals due to pass-alongs (not originally contacted) that purchased prior to the expiration date, and 4) the total number of purchases by individuals (originally contacted or not) after the expiration date.

“Best match” attribution credits a single treatment instance for the response. In the simplest case where the response is a direct exact type response, best match directly credits the treatment instance identified by the response tracking process to be the one treatment instance responsible for the desired behavior. In other cases, for example where the response is a direct inexact type response or an inferred type response which can be attributed to multiple treatment instances, best match will break the tie and credit a single treatment instance based on some configurable criteria (e.g., the treatment instance that promoted to the customer most recently). All other treatment instances receive zero credit under the best match attribution method. For example, if the same promotion is sent to the same individual multiple times over a period of time producing multiple treatment instances, without the means to discern from among the competing treatment instances, best match will credit the treatment instance last sent to the individual. If multiple treatment instances tie based on the specified criteria (e.g., multiple treatment instances were sent on the same latest date for “most recent”), then best match randomly credits any one of the tied treatment instances.

“Fractional attribution” credits one or more treatment instances an equal fraction for the response. Again, in the simplest case where the response is a direct exact type response, fractional attribution directly credits the treatment instance designated by the response tracking process to be the one treatment instance responsible for the desired behavior. Since the direct exact case identifies only one treatment instance, the treatment instance is assigned 1/1 credit. In other cases, again for example where the response is a direct inexact type response or an inferred type response which can be attributed to N treatment instances, fractional attribution will assign each of the N treatment instances 1/N credit (e.g., if there are three treatment instances to receive credit for fractional match, each will receive 1/3 or 0.3333 credit).

“Multiple attribution” assigns multiple treatment instances full credit for the response. For example, where the response is a direct inexact type response or an inferred type response which can be attributed to N treatment instances, multiple attribution will assign each of the N treatment instances full credit (1.0).

“Custom attribution” distributes credit to multiple treatment instances based on a user-defined scheme. Given the set of matching treatment instances and relevant information (e.g., the contact date for each treatment instance, the purchased product, etc.), the custom attribution scheme returns a numeric value representing the credit for each treatment instance. For example, custom attribution can be based on a mathematical formula that credits multiple treatment instances based on exponentially decaying weights assigned from “most recent” to “least recent.” Or given a product hierarchy, a purchased product closest to the promoted product can be given greater credit than a treatment instance where the purchase is less related (e.g., purchasing a TV could credit a promotion targeting TVs more than a promotion targeting electronics in general, or a promotion targeting any product). It is evident that an infinite number of custom attribution possibilities are possible.

22

The response tracking process also can handle different levels of response based on known hierarchies in various dimensions (e.g., customer or product), where each category rolls up fully into a single parent (many-to-one relationships). 5 If the product hierarchy is known and described to the system (e.g., a Panasonic™ DVD player model X belongs to Panasonic™ DVD players which belongs to DVD players, which belongs to Video, which belongs to Electronics, etc.), then a promotion instance promoting the specific model X can trigger different response levels, depending on the actual product purchased. The marketing user can report on the number of responders that purchased exactly model X; the number of responders that did not purchase model X, but purchased a Panasonic™ DVD player; the number of responders that did not purchase a Panasonic™ DVD player, but purchased a DVD player, and so on. Similarly, when a specific individual in a household is targeted, the marketing user wants to know how many times the targeted individual responded vs. how many times another member o the same household responded. 10 In business-to-business scenarios, the customer dimension hierarchy can be significantly more complex (individual, group, department, division, company, etc.).

The response tracking process 40f also tracks control group responses, which are always inferred responses (as the recipients were given no codes to return). As discussed above, hold-out control groups or control groups allow the user to gauge the effectiveness of marketing campaigns targeting actual contact entities. Control groups gauge whether entities of a control group perform the desired behavior (e.g., inquire, 15 activate, purchase) on a desired product (e.g., money-market account, plasma television, Boston-Los Angeles flight) without having been promoted the product/service. In this regard, response tracking for control groups is similar to response tracking for inferred responses, as control group response tracking is performed only of the promotion attributes, and response codes are ignored, if any. That is, the response codes reported with the response correspond to treatment instances of promotion version actually promoted to contacts and do not correspond to treatment instances of the control group.

After attributing response to treatment instances 436, if any response codes were provided 438, those response codes are removed 440. Control treatment instances containing the responder (i.e., where the responder was in the control group) are identified by matching promotion attributes 442. If the 20 response date is between the effective and expiration dates of any control treatment instances 444, then discard any control treatment instances in which the response was not within the valid date range 446. The identified treatment instances of the control groups can then be credited and the appropriate response attributes recorded 448. Control group response tracking can be accomplished with, but is not limited to, multiple attribution. After control treatment instances have been credited, the response tracking process 40f returns to process the next event in the action table 420.

In addition to direct and inferred responses, the response tracking process can also track personalized promotions through data-driven parameterization and its associated controls. As described above, data-driven promotion personalization refers to the case where contact entities of a single target group receive promotion versions based on parameterized promotion attributes with values populated from a list or table containing personalized data for each of the contact entities. For the data in personalized case to be meaningful for analysis, the response reports information pertaining specifically to the data-driven parameterized promotion attributes. The detailed contact history table records the exact promotion version(s) received by each member of the target group (or the 25 30 35 40 45 50 55 60 65

US 8,315,904 B2

23

offer versions a member of the control group would have received). The response tracking process 40/ is then the same as previously described, except that for each treatment in the treatment history table 256 (FIG. 15), a flag identifies whether corresponding contact history data resides in the contact history table 258 or the detailed contact history table (not shown in FIG. 15).

In FIG. 20, the action table can be used to produce one or more response tables 414 with rows to log the response attribution data and the response attributes data. A response table is (conceptually) produced for each audience level. Therefore, for example, there can be a response table for the household audience as identified by the householdID, one for the customer level as identified by the custID, and one for the account level as identified by the accountID (note that these could be same underlying physical table if desired). The response table permits the user to review the data from the marketing campaign to assess campaign performance based on, for example, return on investment or any other economic, business, or investment model. To assist the user in this regard, the response tracking process can produce varied response tracking reports. As examples, reports can include detailed response type information. Reports can be generated according to time, group, campaign, channel, promotion instance, or promotion version attributes. Reports can be group-based (i.e., reporting the number of responses in view of the number of groups in a campaign or any group-level custom attribute) or promotion-based (i.e., reporting based on the different promotions within a campaign or any promotion custom attribute). Reports can also show the effectiveness of a given promotion by the number of targeted responders in view of the number of viral (non-targeted) responders. Reports can also show control group data and can calculate lift over the control group. Reports can be configured to display unique vs. duplicate counts. In some implementations, reports include data from any combination of the above.

Processors suitable for the execution of a computer program include, by way of example, both general and special purpose microprocessors, and any one or more processors of any kind of digital computer. Generally, a processor will receive instructions and data from a read-only memory or a random-access memory or both. The essential elements of a computer are a processor for executing instructions and one or more memory devices for storing instructions and data. Generally, a computer will also include, or be operatively coupled to receive data from or transfer data to, or both, one or more mass storage devices for storing data, e.g., magnetic, magneto optical disks, or optical disks. Information carriers suitable for embodying computer program instructions and data include all forms of non-volatile memory, including by way of example semiconductor memory devices, e.g., EPROM, EEPROM, and flash memory devices; magnetic disks, e.g., internal hard disks or removable disks; magneto optical disks; and CD ROM and DVD-ROM disks. The processor and the memory can be supplemented by, or incorporated in special purpose logic circuitry.

To provide for interaction with a user, embodiments can be implemented on a computer having a display device, e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor, for displaying information to the user and a keyboard and a pointing device, e.g., a mouse or a trackball, by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback, e.g., visual feedback, auditory feedback, or tactile feedback; and input from the user can be received in any form, including acoustic, speech, or tactile input.

24

Embodiments can be implemented in a computing system that includes a back end component, e.g., as a data server, or that includes a middleware component, e.g., an application server, or that includes a front end component, e.g., a client computer having a graphical user interface or a Web browser through which a user can interact with an implementation of embodiments, or any combination of such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication, e.g., a communication network. Examples of communication networks include a local area network (LAN) and a wide area network (WAN), e.g., the Internet.

The system and method may use the "World Wide Web" (Web or WWW), which is that collection of servers on the Internet that utilize the Hypertext Transfer Protocol (HTTP). HTTP is a known application protocol that provides users access to resources, which may be information in different formats such as text, graphics, images, sound, video, Hyper-text Markup Language (HTML), as well as programs. Upon specification of a link by the user, the client computer makes a TCP/IP request to a Web server and receives information, which may be another Web page that is formatted according to HTML. Users can also access other pages on the same or other servers by following instructions on the screen, entering certain data, or clicking on selected icons. It should also be noted that any type of selection device known to those skilled in the art, such as check boxes, drop-down boxes, and the like, may be used for embodiments using web pages to allow a user to select options for a given component. Servers run on a variety of platforms, including UNIX machines, although other platforms, such as Windows 2000/2003, Windows NT, Sun, Linux, and Macintosh may also be used. Computer users can view information available on servers or networks on the Web through the use of browsing software, such as Netscape Navigator, Microsoft Internet Explorer, Mosaic, or Lynx browsers. The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

A number of embodiments have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the application. Other embodiments are within the scope of the following claims.

What is claimed is:

1. A computer implemented method comprising:
producing, by one or more computers, a promotion list for a promotion management campaign by:
generating, by one or more computers, a promotion instance from a promotion template;
receiving, by one or more computers executing marketing campaign software, a search query that includes one or more attributes of a promotion instance;
searching one or more data repositories for promotion instances having attributes corresponding to the attributes specified in the search query;
returning a list including one or more promotion instances having the attributes corresponding to the attributes specified in the search query;
receiving, by the one or more computers, a selection of one or more promotion instances, from the returned list, to be included in the promotion list;
assigning the selected promotion instances to the promotions list; and
storing the promotion list in an electronic medium.

US 8,315,904 B2

25

2. The method of claim 1 wherein assigning the selected promotion instances to the promotions list comprises:

querying promotion instances to find one or more matching promotion instances having matching attributes to one or more of the promotion attributes; and at runtime, adding the matching promotion instances to the promotions list.

3. The method of claim 1 wherein assigning the selected promotion instances to the promotions list comprises:

displaying the one or more promotion instances stored in the electronic medium in a searchable and browsable view; and adding from the searchable and browsable view, one or more promotion instances to the promotions list.

4. The method of claim 2 further comprises:

limiting the matching promotion instances to an N number of promotions, wherein N is an integer.

5. The method of claim 2 further comprises:

sorting the matching promotion instances in one of an ascending or descending order based on one or more promotion attributes.

6. The method of claim 2 further comprises:

filtering the matching promotion instances to matching promotion instances located in one or more folders of the electronic medium.

7. The method of claim 1 further comprises:

assigning to the promotion list a security policy governing which users are allowed to interact with the promotion list.

8. The method of claim 1 wherein the promotion list includes promotion attributes, the promotion attributes comprising:

a promotion list name;
an icon; and
a promotion list description.

9. The method of claim 2 wherein a query comprises clauses including one or more of including a function, a variable, a promotion attribute connected by Boolean operators to match to one or more attributes of one or more promotions.

10. The method of claim 1 wherein the one or more promotion instances comprise a first promotion instance having at least one or more different promotion attributes or different promotion attribute values than a second promotion instance.

11. A computer program product residing on a computer readable medium comprises instructions for causing a computer to:

produce a promotion list for a promotion management campaign by comprising instructions for causing a computer to:

generate a promotion instance from a promotion template;

receive a search query that includes one or more attributes of a promotion instance;

search one or more data repositories for promotion instances having attributes corresponding to the attributes specified in the search query;

return a list including one or more promotion instances having the attributes corresponding to the attributes specified in the search query;

receive a selection of one or more promotion instances, from the returned list, to be included in the promotion list;

assign the selected promotion instances to the promotions list; and

store the promotion list in an electronic medium.

26

12. The computer program product of claim 11 wherein instructions that assign the selected promotion instances to the promotions list further comprise instructions for causing a computer to:

query promotion instances to find one or more matching promotion instances having matching attributes to one or more of the promotion attributes; and at runtime, add the matching promotion instances to the promotions list.

13. The computer program product of claim 11 wherein instructions that assign the selected promotion instances to the promotions list further comprise instructions for causing a computer to:

display the one or more promotion instances stored in the electronic medium in a searchable and browsable view; and add from the searchable and browsable view, one or more promotion instances to the promotions list.

14. The computer program product of claim 12 further comprises instructions for causing a computer to:

limit the matching promotion instances to an N number of promotions, wherein N is an integer.

15. The computer program product of claim 12 further comprises instructions for causing a computer to:

sort the matching promotion instances in one of an ascending or descending order based on one or more promotion attributes.

16. The computer program product of claim 12 further comprises instructions for causing a computer to:

filter the matching promotion instances to matching promotion instances located in one or more folders of the electronic medium.

17. The computer program product of claim 11 further comprises instructions for causing a computer to:

assign to the promotion list a security policy governing which users are allowed to interact with the promotion list.

18. The computer program product of claim 11 wherein the promotion list includes promotion attributes, the promotion attributes comprising:

a promotion list name;
an icon; and
a promotion list description.

19. The computer program product of claim 12 wherein a query comprises clauses including one or more of including a function, a variable, a promotion attribute connected by Boolean operators to match to one or more attributes of one or more promotions.

20. The computer program product of claim 11 wherein the one or more promotion instances comprise a first promotion instance having at least one or more different promotion attributes or different promotion attribute values than a second promotion instance.

21. An apparatus comprises:

a processor;
memory executing on one or more computers a computer program product comprising instructions for causing the processor to:

produce a promotion list for a promotion management campaign by:

generating a promotion instance from a promotion template;

receiving a search query that includes one or more attributes of a promotion instance;

searching one or more data repositories for promotion instances having attributes corresponding to the attributes specified in the search query;

US 8,315,904 B2

27

returning a list including one or more promotion instances having the attributes corresponding to the attributes specified in the search query;
receiving a selection of one or more promotion instances, from the returned list, to be included in the promotion list;

5

28

assigning the selected promotion instances to the promotions list; and
storing the promotion list in an electronic medium.

* * * * *

EXHIBIT D



US007702719B1

(12) **United States Patent**
Betz et al.

(10) **Patent No.:** US 7,702,719 B1
(45) **Date of Patent:** Apr. 20, 2010

(54) **METHODS AND APPARATUS FOR REDUCING THE NUMBER OF SERVER INTERACTIONS IN NETWORK-BASED APPLICATIONS USING A DUAL-MVC APPROACH**

(75) Inventors: **Katherine Betz**, Hopewell Junction, NY (US); **Avraham Leff**, New Hempstead, NY (US); **James Thomas Rayfield**, Ridgefield, CT (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 883 days.

(21) Appl. No.: **09/500,208**

(22) Filed: **Feb. 8, 2000**

(51) **Int. Cl.**
G06F 15/16 (2006.01)

(52) **U.S. Cl.** **709/203**; 709/201; 709/205;
709/217; 709/223

(58) **Field of Classification Search** 709/201–203,
709/205, 217–219, 223, 224; 715/733, 747
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,572,643 A *	11/1996	Judson	709/218
5,768,510 A *	6/1998	Gish	709/218
5,838,906 A *	11/1998	Doyle et al.	709/202
5,926,177 A *	7/1999	Hatanaka et al.	715/747
6,003,047 A *	12/1999	Osmond et al.	715/513
6,035,332 A *	3/2000	Ingrassia et al.	709/218
6,151,621 A *	11/2000	Colyer et al.	709/217
6,161,136 A *	12/2000	Hyndman et al.	709/223
6,170,019 B1 *	1/2001	Dresel et al.	709/330
6,272,556 B1 *	8/2001	Gish	709/315

6,356,933 B2 *	3/2002	Mitchell et al.	709/203
6,393,569 B1 *	5/2002	Orensteijn	709/203
6,437,803 B1 *	8/2002	Panasyuk et al.	715/733
6,496,202 B1 *	12/2002	Prinzing	715/762
6,505,246 B1 *	1/2003	Land et al.	709/224
6,950,850 B1 *	9/2005	Leff et al.	709/203
6,968,534 B1 *	11/2005	Hayase	717/105

FOREIGN PATENT DOCUMENTS

JP	10-207805	8/1998
JP	10-303893	11/1998
JP	10-320390	12/1998
JP	11-003237	1/1999
JP	11-031145	2/1999
JP	11-031155	2/1999
JP	11-338888	12/1999

OTHER PUBLICATIONS

G.E. Krasner and S.T. Pope, "A Cookbook for Using the Model-View-Controller User Interface Paradigm in SmallTalk-80," Journal of Object-Oriented Programming, 1(3):26-49, Aug./ Sep. 1988.

* cited by examiner

Primary Examiner—Ramy Mohamed Osman

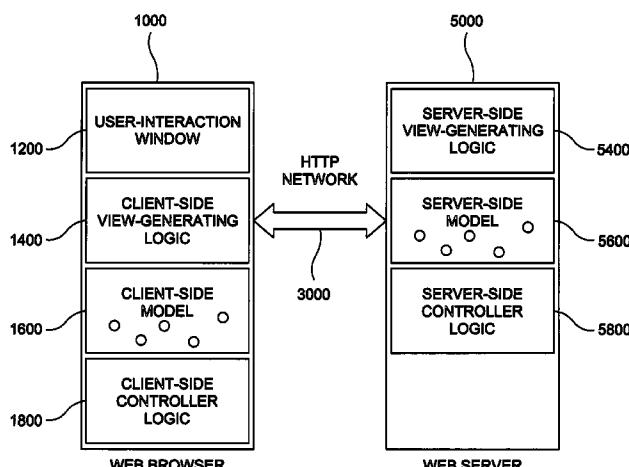
(74) Attorney, Agent, or Firm—Kenneth R. Corsello; Ryan, Mason & Lewis, LLP

(57)

ABSTRACT

In accordance with a dual-MVC (Model-View-Controller) approach of the invention, the client machine maintains part of the Model, eliminating many round-trips to the server. The invention provides an implementation approach for constructing the frame(s) contained in the client's browser software to allow it to be refreshed locally. The invention addresses development and maintenance by specifying a library of functions that a programmer can use to construct the View and Controller on the client machine. When part of the Model is maintained on the client, the client has the ability to respond to user interaction without paying the performance penalty of continually interacting with the server.

26 Claims, 12 Drawing Sheets



U.S. Patent

Apr. 20, 2010

Sheet 1 of 12

US 7,702,719 B1

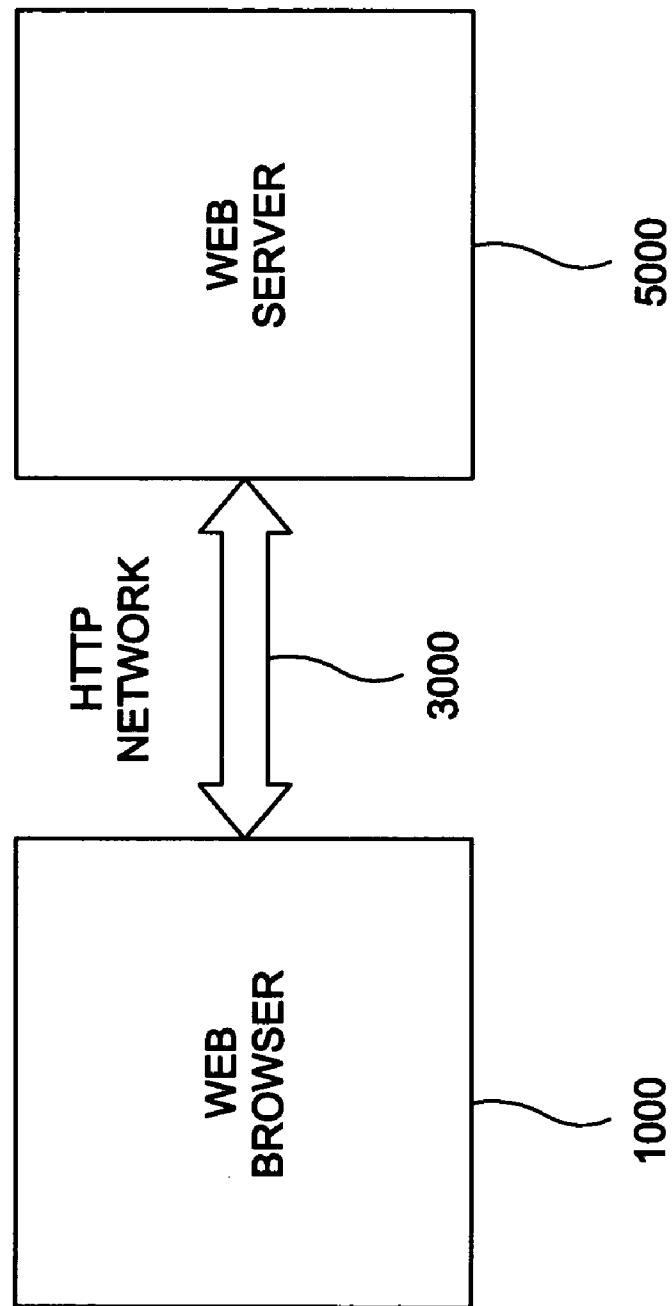


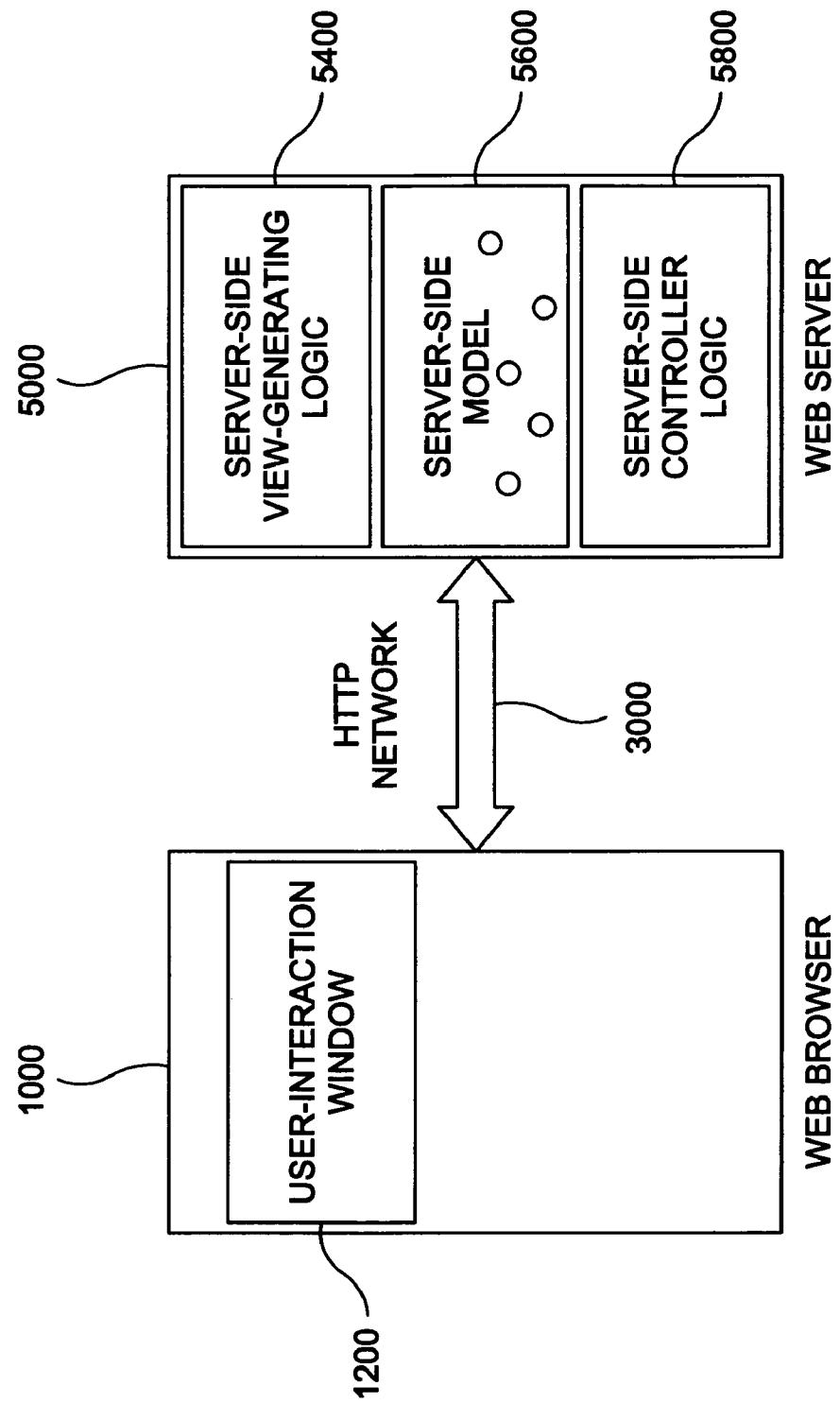
FIG. 1

U.S. Patent

Apr. 20, 2010

Sheet 2 of 12

US 7,702,719 B1

FIG. 2 PRIOR ART

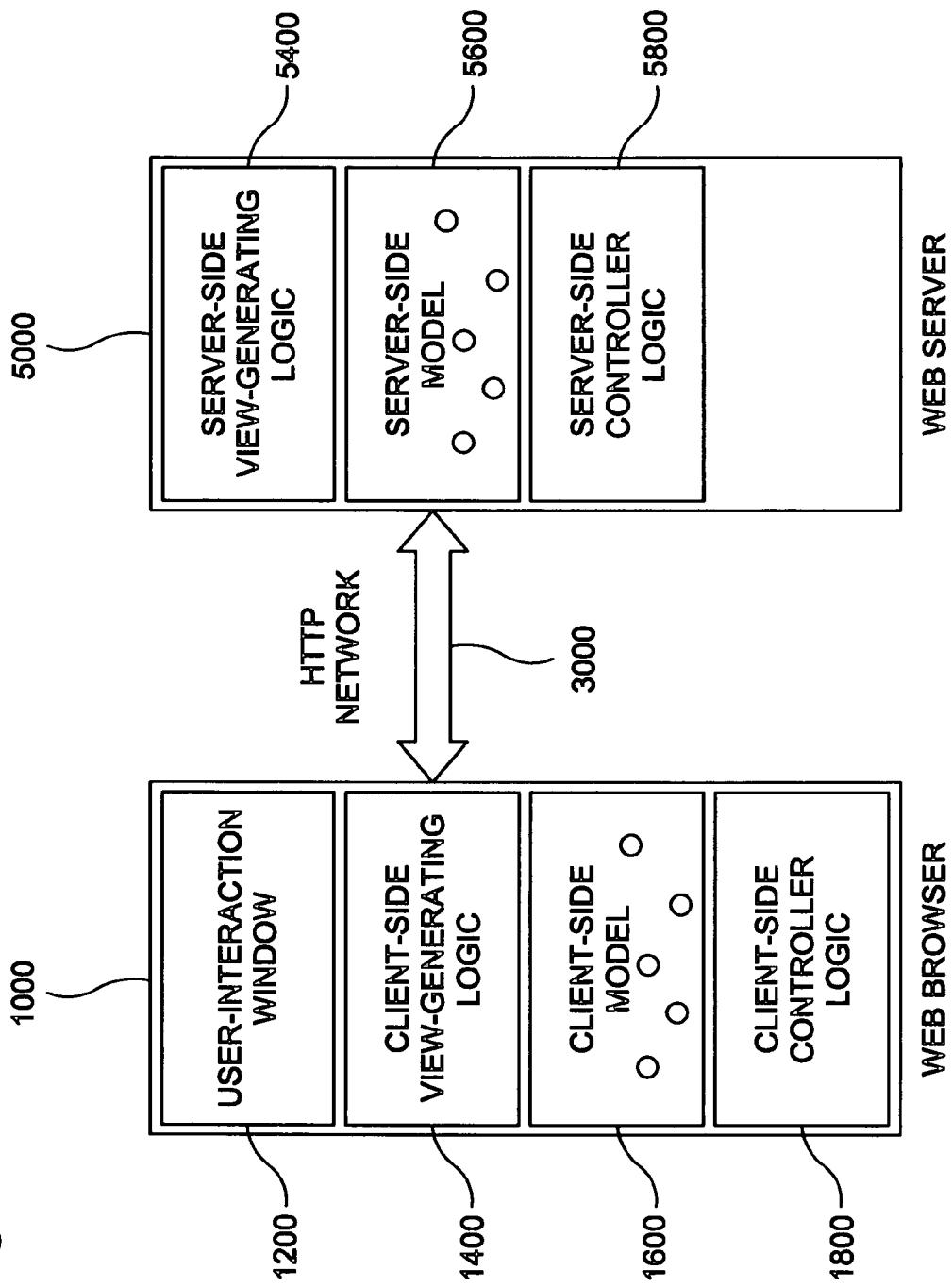
U.S. Patent

Apr. 20, 2010

Sheet 3 of 12

US 7,702,719 B1

FIG. 3



U.S. Patent

Apr. 20, 2010

Sheet 4 of 12

US 7,702,719 B1**FIG. 4**

ID	NAME	PHONE	OPERATING SYSTEM	REMARKS
1234567	John Smith	(914) 222-7890	NT	

ADD ANOTHER PROBLEM**SAVE**

U.S. Patent

Apr. 20, 2010

Sheet 5 of 12

US 7,702,719 B1

FIG. 5

ID	NAME	PHONE	OPERATING SYSTEM	REMARKS
1234567	John Smith	(914) 222-7890	NT <input checked="" type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/>

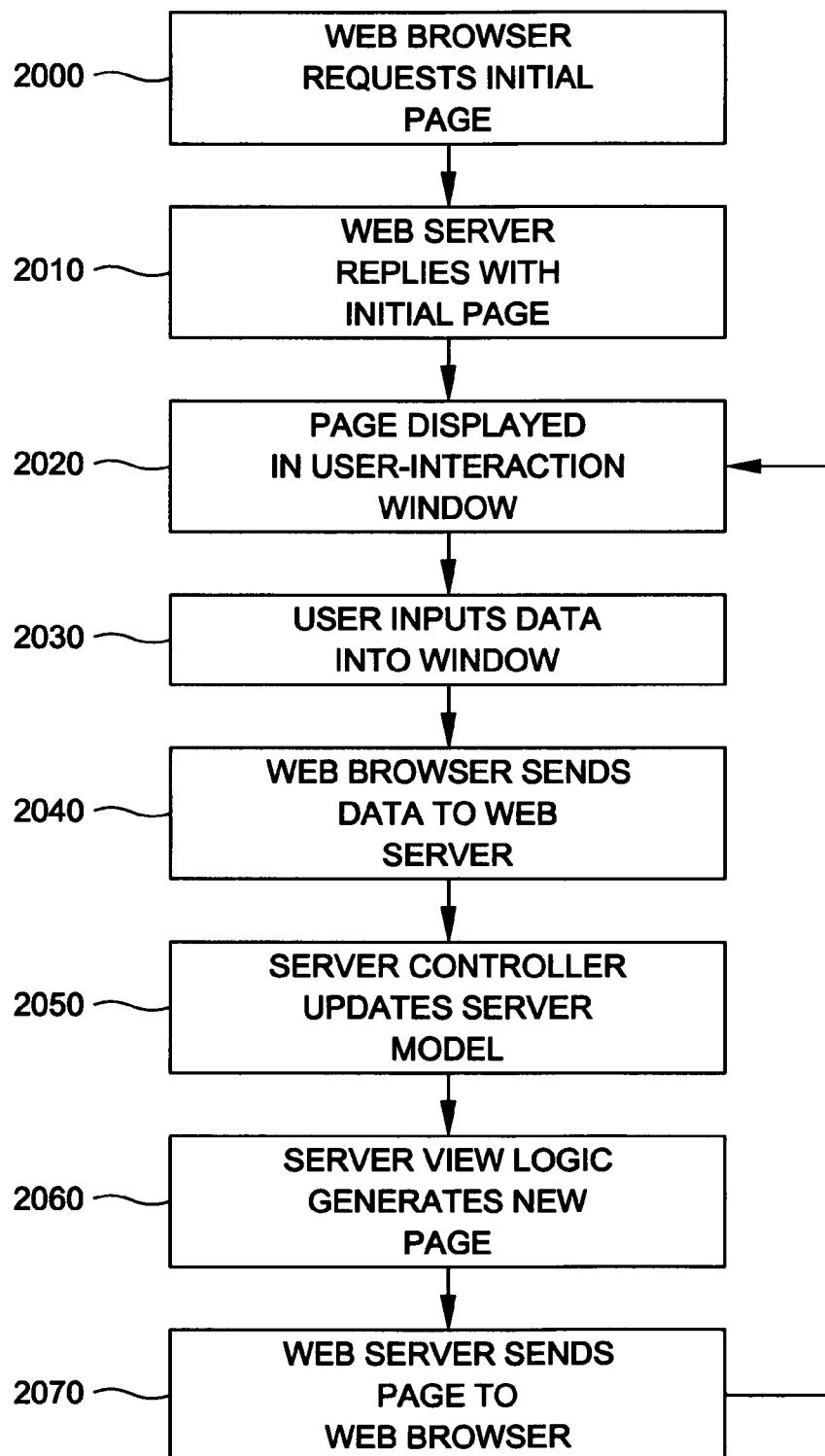
VERSION	PLATFORM	HOST NAME	IP ADDRESS

U.S. Patent

Apr. 20, 2010

Sheet 6 of 12

US 7,702,719 B1

FIG. 6 PRIOR ART

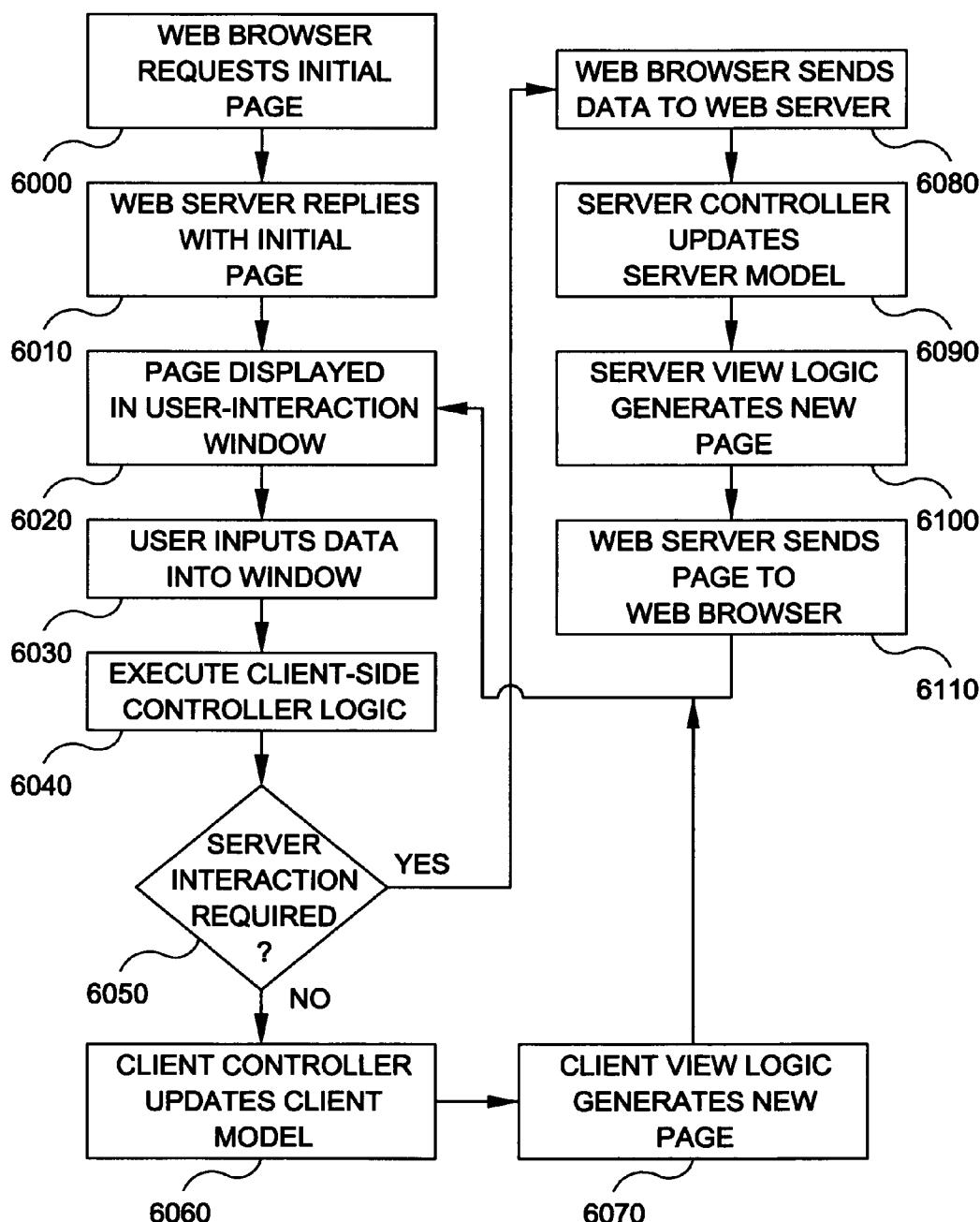
U.S. Patent

Apr. 20, 2010

Sheet 7 of 12

US 7,702,719 B1

FIG. 7

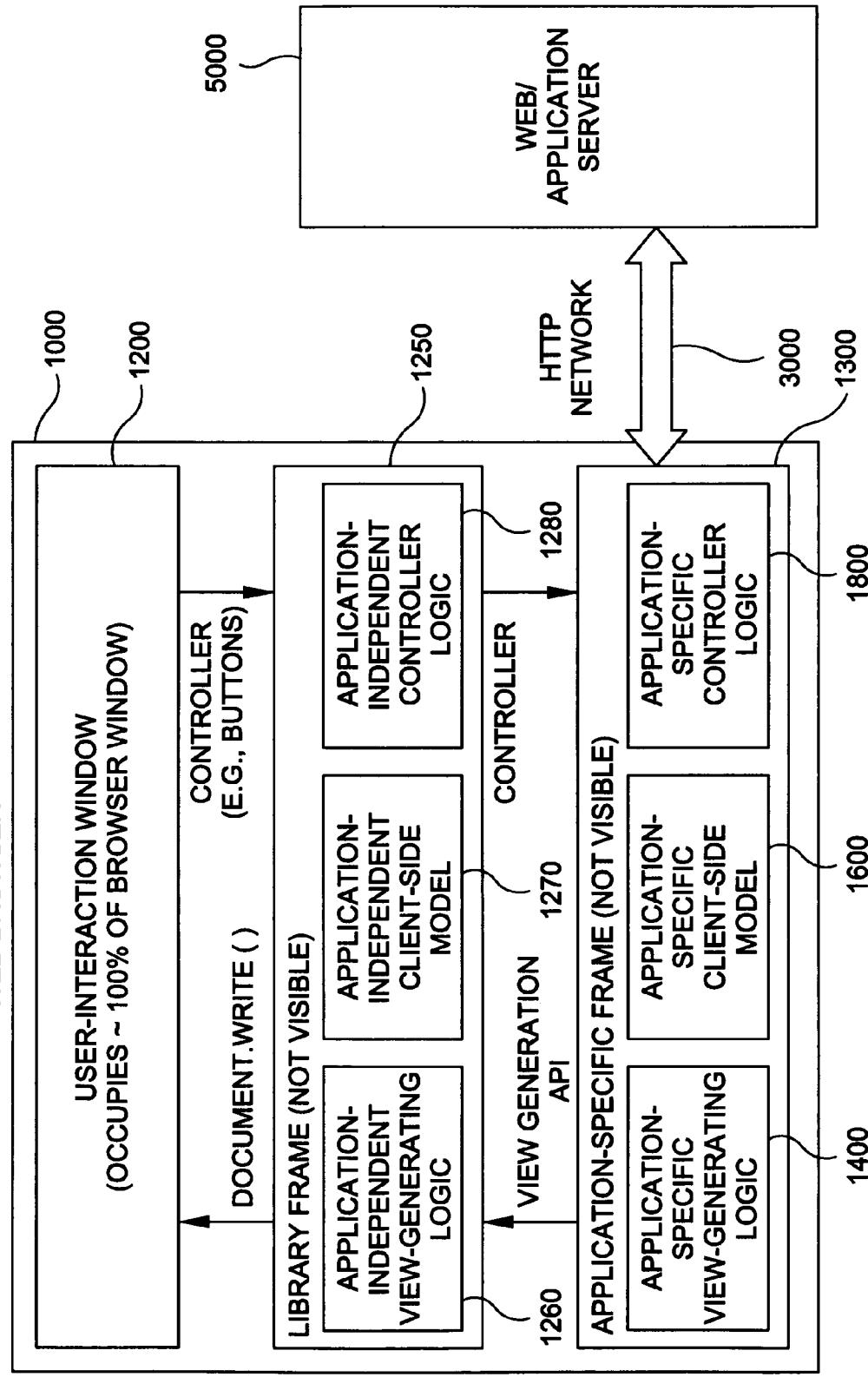


U.S. Patent

Apr. 20, 2010

Sheet 8 of 12

US 7,702,719 B1

FIG. 8

U.S. Patent

Apr. 20, 2010

Sheet 9 of 12

US 7,702,719 B1

FIG. 9A

```
<SCRIPT LANGUAGE="JavaScript">

// Define the variable (s) holding the Model data
var counter = 0; // a counter value

// Define the variables holding pointers to the View components
var theFrame; // pointer to the LFrame component
var theForm; // pointer to the LForm component
var theText; // pointer to the LText component
var theButton; // pointer to the LButton component
var theStaticText; // pointer to the LStaticText component

function initialize () // Create and initialize components
{
    // Create a library LFrame component
    theFrame = new parent.LIBFrame.LFrame ("Frame1", "A Frame");

    // Adjust the base font to be +3 larger
    theFrame.adjustBaseFont (+3);

    // Create a library LForm component inside the LFrame
    theForm = new parent.LIBFrame.LForm ("Form1", theFrame);

    // Create a library LText (text entry field) component inside the LForm
    theText = new parent.LIBFrame.LText ("foo", "bar", theForm);

    // Set up an event handler to handle user changes to the LText
    theText.setOnChange (T_onChange);

    //Create a library LButton component inside the LForm
    theButton = new parent.LIBFrame.LButton ("increment_button",
                                           "Increment Counter", theForm);

    // Set an event handler to handle clicks on the LButton
    theButton.setonClick (B_increment);
}
```

(CONTINUED IN FIG. 9B)

U.S. Patent

Apr. 20, 2010

Sheet 10 of 12

US 7,702,719 B1

FIG. 9B

(CONTINUED FROM FIG. 9A)

```
// Create a library LStaticText (label) component in the LForm  
theStaticText = new parent.LIBFrame.LStaticText ("OST", "piece of text ...",  
    theForm);  
  
// Set the font size for the LStaticText  
theStaticText.setFontSize (1);  
  
//Set the color for the LStaticText  
theStaticText.setColor ("red");  
  
// Refresh the contents of the User Interaction window  
theFrame.RefreshVisibleFrame ();  
  
// Set the browser focus to be on the LText component  
theText.focus ();  
}  
  
// This function is invoked on the LText component when the value is changed  
// by a user interaction  
function T_onChange ()  
{  
    // Change the value to a message  
    this.setValue ("Please do not change the value!");  
}  
  
// This function is invoked on the LText object when the user clicks on it
```

(CONTINUED IN FIG. 9C)

U.S. Patent

Apr. 20, 2010

Sheet 11 of 12

US 7,702,719 B1

FIG. 9C

(CONTINUED FROM FIG. 9B)

```
function T_onClick ()
{
    this.setValue ("Please do not click on this!");
}

// This function is invoked on the LButton object when the user clicks on it
function B_increment ()
{
    // Increment the counter (Model)
    counter += 1;

    // Move the new counter value into the LText field
    theText.setValue (counter);

    // Increment the font size on the LStaticText component
    theStaticText.setFontSize (theStaticText.getFontSize() + 1);

    // Toggle the LStaticText component font between Courier and Impact
    if (0 == (counter % 2))
        theStaticText.setFace ("Courier");
    else
        theStaticText.setFace ("Impact");
}

</SCRIPT>
```

U.S. Patent

Apr. 20, 2010

Sheet 12 of 12

US 7,702,719 B1

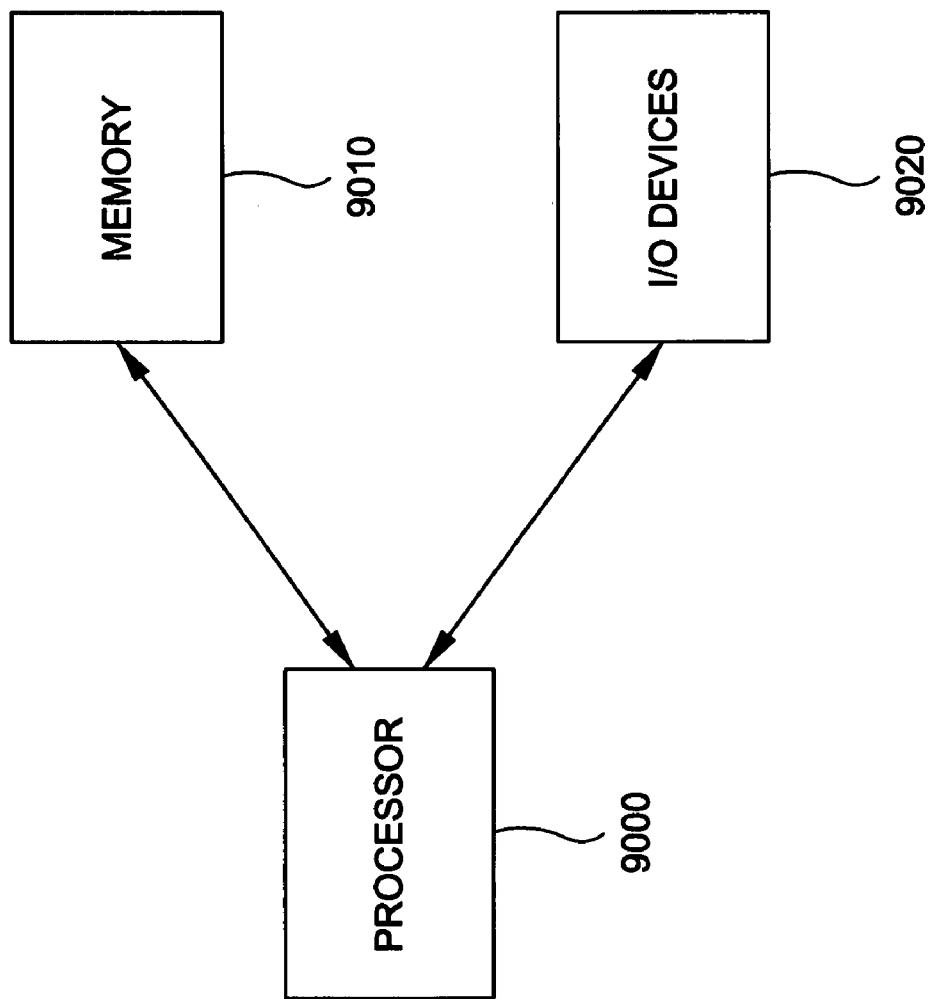


FIG. 10

US 7,702,719 B1

1

**METHODS AND APPARATUS FOR
REDUCING THE NUMBER OF SERVER
INTERACTIONS IN NETWORK-BASED
APPLICATIONS USING A DUAL-MVC
APPROACH**

FIELD OF THE INVENTION

The present invention is related to network-based client/server applications and, more particularly, to methods and apparatus for reducing the number of server interactions in world wide web-based applications using a dual Model-View-Controller approach.

BACKGROUND OF THE INVENTION

Typical world wide web (e.g., Internet/Intranet) applications allow users to access and update data on remote servers. The remote server contains the master application data and the client displays views of this data. Examples of such applications include a help desk, life insurance underwriting, health insurance pre-certification for medical procedures, and automated teller machines.

Historically (circa 1970), applications executed on large mainframe systems (servers), to which clients connected via "dumb terminals," e.g., Teletypes, DEC VT100, IBM 327x-series terminals, etc. In these architectures, very little processing was done on the client side. Subsequently (circa 1980), as hardware became cheaper, some of the processing was moved to the client-side hardware, in so-called "fat-client" architectures. A fat client maintains some of the application state, enabling some processing to occur locally thus eliminating round-trips to the server.

Fat-client applications can be described in terms of the MVC (Model-View-Controller) paradigm. The MVC paradigm is described, for example, in G. E. Krasner and S. T. Pope, "A Cookbook for Using the Model-View-Controller User Interface Paradigm in SmallTalk-80," *Journal of Object-Oriented Programming*, 1(3):26-49, August/September 1988, the disclosure of which is incorporated herein by reference. In this paradigm, the "Model" contains the data, rules, and algorithms affecting the data. The "View" is a screen or window representation of a subset of the model that the application chooses to display. The "Controller" is the logic that processes user requests, such as pressing a button. The Controller causes the Model to be changed and/or the View to be refreshed.

Fat-client applications may maintain a Model and Controller on the server but never maintain a View or View Generation Logic (VGL) on the server. Instead, View and VGL are maintained solely by the fat client.

While fat-client applications provided improved response time, the distribution and maintenance of the client software and databases was problematic. Users tended to customize their client system, and this made it difficult to develop a client-side software base that would work properly on all systems. In addition, when the client software needed to be updated, a company had to figure out how to update all the client machines in the field.

Applets (see <http://java.sun.com/applets/index.html>) are one implementation of the fat-client architecture. When the user starts the application, or while running the application, the server downloads part of the application to the client's machine to execute locally. This improves performance because round-trips to the server are lessened. However, when the applet is first invoked, it must be downloaded to the client, which in practice can take a significant length of time.

2

Also, in practice, web browsers often support different levels of Java Virtual Machines (JVMs). Because an applet is interpreted by the JVM, the difference in JVMs causes inconsistent operation of applets on different browsers or even different versions of the same browser.

More recently (circa 1994), the development and popularization of the world wide web has led towards a so-called "thin-client" application architecture. In this architecture, most of the application logic again executes on the server, with only the browser display logic, i.e., application-independent display logic, executing inside the client-side web browser software. Here the browser is the client, and displays a View. Each time the user interacts with the View, the remote server is notified to update the View and/or Model.

Another implementation of client/server interaction over the web uses Dynamic HyperText Markup Language (DHTML). DHTML is described, for example, in D. Goodman, "Dynamic HTML, The Definitive Reference," O'Reilly, 1998, the disclosure of which is incorporated herein by reference. DHTML is a superset of HTML (HyperText Markup Language) which includes JavaScript and Cascading Style Sheets (CSS). JavaScript can be used to implement much of the same functionality that applets provide, is in practice more portable between browsers, and is quicker to download.

Although fat-client architectures provide the most flexibility and performance, their maintenance and distribution problems, together with the popularity of the web, have led many application developers to adopt thin-client architectures for new applications. A problem with this architecture is that network communication delays during server interactions again lead to reduced performance.

Therefore, a need exists for a new Model-View-Controller architecture for Internet/Intranet applications which does not require continual network communication between the client and server.

SUMMARY OF THE INVENTION

The invention relates to web applications and, more particularly, to an architecture and method of programming which improves performance and eases the development and maintenance effort. The invention addresses performance by employing a dual-MVC approach, in which a subset of the application's Model-View-Controller reside on the client, and the full Model-View-Controller and View-Generating-Logic reside on the server, thereby reducing the number of required server interactions. In contrast to the fat-client architecture, the invention does not require any application installation or persistence on the client.

More specifically, the classic fat-client implementation requires that the user (or administrator) physically install the client software on the client machine, and the application will require the long-term use of persistent storage (i.e., disk space) on the client machine. For example, consider the well-known tax preparation assistance software program known as "TurboTax" as a fat-client application. A CDROM is required to run the install process. After installation, the program uses a sizable amount of disk space, both for its own program/data files and for the user's tax return. From time to time, e.g., during electronic filing of a tax return, "TurboTax" interacts with a remote server as a (fat) client.

The applet approach avoids some of the problems of fat clients, because the applet is downloaded automatically and does not make long-term use of disk space. However, the applet may take a long time to download, and it is difficult in practice to write an applet which will function correctly on every client.

US 7,702,719 B1

3

The dual-MVC approach of the invention avoids the installation and disk-space problems because client-side code associated with the MVC of the application is preferably downloaded from the server on demand by the user of the client device, avoiding the install process, and is preferably not saved permanently to the disk of the client. The dual-MVC approach also avoids problems associated with applets because the MVC code downloaded according to the invention is physically smaller than an applet and thus results in a faster download. Also, the dual-MVC approach of the invention preferably uses HTML and JavaScript constructs which are, in practice, more portable between browsers than are applets.

Particularly, in accordance with the unique dual-MVC approach of the invention, the client machine maintains part of the Model, eliminating many round-trips to the server. The invention provides an implementation approach for constructing the frame(s) contained in the client's browser software to allow it to be refreshed locally. The invention addresses development and maintenance by specifying a library of functions that a programmer can use to construct the View and Controller on the client machine. When part of the Model is maintained on the client, the client has the ability to respond to user interaction without paying the performance penalty of continually interacting with the server.

It is to be appreciated that the term "frame" as used herein has a similar usage as in HTML. Web browsers typically display one or more windows on the client's screen. Each window may correspond to one HTML frame, or may instead correspond to one HTML "frameset." A frameset comprises one or more frames, which are like sub-windows inside a frameset window.

The invention specifies a way to structure browser-based clients so that classic MVC interactions are possible. The client's browser screen may be divided into multiple frames. In the case of a web browser running at the client, the web browser includes an HTML frameset. There may be one or more visible frames that the user interacts with, and a number of "invisible" frames where the logic resides. By "invisible" frame, we mean that the frames height or width has been set to a very small value, so that they appear not to occupy any screen area. Visible frames obviously occupy some screen area. The invisible frames contain the application logic and the logic to rewrite the visible frames. It is to be appreciated that an application may create additional windows or framesets, and/or have additional frames within a window, where the window corresponds to a frameset. Without this invention, a client-side application attempting to update its View will destroy itself by overwriting its own Controller logic and Model data.

It is to be appreciated that the term "logic" as used herein is intended to refer to program code, preferably in JavaScript at the client side, although it may also be in Java, VBScript, C++, C, or any other programming language which is supported by the browser.

The invention also specifies a way to program this kind of web application. Programmers using only the dual-MVC and multiple frames will need to generate HTML within their View-generating logic. This requires the programmer to think in two different domains simultaneously: JavaScript for the View-generating logic, Model, and Controller, and HTML for the View itself. This may be a complicated, tedious, and error-prone task depending on the application being designed. To substantially eliminate this situation, the present invention specifies a library of functions that are capable of writing the HTML with JavaScript. The programmer can implement the application using these functions rather than

4

having to write the complicated HTML/JavaScript. This is analogous to programmers writing user interfaces using the Java Swing library.

These and other objects, features and advantages of the present invention will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram illustrating an exemplary client/server system suitable for employing the present invention;

FIG. 2 is a block diagram illustrating a conventional thin-client architecture;

FIG. 3 is a block diagram illustrating a dual-MVC architecture according to an embodiment of the present invention;

FIG. 4 is a diagram illustrating an initial display screen associated with an exemplary application;

FIG. 5 is a diagram illustrating a second display screen associated with an exemplary application;

FIG. 6 is a flow diagram illustrating application flow for a conventional thin-client architecture;

FIG. 7 is a flow diagram illustrating application flow for a dual-MVC architecture according to an embodiment of the present invention;

FIG. 8 is a block diagram illustrating a client-side frameset architecture for a dual-MVC approach according to an embodiment of the present invention;

FIGS. 9A through 9C depict code fragments illustrating a sample application using a library frame according to an embodiment of the present invention; and

FIG. 10 is a block diagram illustrating a hardware implementation of a client computer system and/or server computer system suitable for implementing the dual-MVC methodologies of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The following description will illustrate the invention using an exemplary web-based client/server system. It should be understood, however, that the invention is not limited to use with any particular client/server system. The invention is instead more generally applicable to any network-based client/server system in which it is desirable to substantially reduce server interactions on the part of the client when executing one or more applications associated with the server.

FIG. 1 is a block diagram illustrating an exemplary client/server system suitable for employing the present invention. The system comprises a client computer system which is configured to run web browser software 1000. It is to be appreciated that any suitable web browser software may be employed and that the methodologies of the invention are not limited to any particular web browser software. Further, as shown in the client/server system of FIG. 1, the web browser

1000 communicates over a network 3000 with a web server 5000. The network 3000 may employ a HyperText Transport Protocol (HTTP), as is well known in the art and as is implemented in accordance with the world wide web, however, other suitable protocols and networks may be employed. For example, the client and server may alternatively be connected via a private network, a local area network, or some other suitable network.

It is to be appreciated that an example of such a client/server system as illustrated in FIG. 1 may include an arrangement wherein a user at a personal computer or workstation

US 7,702,719 B1

5

(client) accesses a particular application (e.g., help desk, life insurance underwriting, health insurance pre-certification for medical procedures, and automated teller machines) associated with a particular server on the HTTP network in accordance with web browser software loaded and running on the client machine.

FIG. 2 illustrates a block diagram of a conventional thin-client architecture as may be implemented in a web browser 1000 and server 5000. In this architecture, the application's view-generating logic 5400, Model 5600, and Controller 5800 reside entirely on the server 5000. The client's browser contains the user-interaction window 1200.

In contrast, referring now to FIG. 3, a block diagram illustrating a dual-MVC architecture according to an embodiment of the present invention is shown. In this architecture, the web server 5000 comprises Model-View-Controller components, including server-side View-generating logic 5400, server-side Model data 5600 and server-side Controller logic 5800. The web browser (client) 1000 comprises a user-interaction window 1200, client-side View-generating logic 1400, client-side Model data 1600 and client-side Controller logic 1800. These MVC components are preferably downloaded to the client from the server at the time that the client requests use of the application at the server. Given the fact that the client and server both contain MVC components, we term this approach dual-MVC.

A principal advantage of our invention is a dramatic improvement in application response time, since many interactions with the server are eliminated. To demonstrate the performance improvement which may be achieved, a help desk application will be described. In accordance with such a help desk application, a computer user (i.e., customer) may have problems with his/her computer and call technical support. The person providing the help (i.e., consultant) has a web application to record information about the customer and his/her problem. In other words, a help desk consultant sits in front of a web-based user interface and uses the interface to communicate with the help desk server in order to record and track a customer's problems which he/she reports during a call to the consultant. In such an example, the computer system with the user interface that the help desk consultant is using is considered the client. Specifically, the consultant utilizes a web browser 1000 (FIG. 3) running on the client to access the help desk server. The help desk server is the server 5000 (FIG. 3). FIG. 4 illustrates an example of the consultant's initial screen, corresponding to the user-interaction window 1200 of FIG. 3. Depending on the operating system selected, a new table is added to the screen to collect operating specific information about the customer's computer. This second screen with the new table is shown in FIG. 5.

We first describe the application flow for an implementation using the conventional thin-client architecture as shown in FIG. 2. FIG. 6 shows a generic application flow for the conventional thin-client architecture. In step 2000, the web browser 1000 requests the initial page from the web server 5000, using the network 3000. In step 2010, the web server replies with the initial page and, in step 2020, the web browser displays the page in the user-interaction window 1200. The user interacts with the displayed page, in step 2030, by inputting data into the window, and the web browser transmits the data back to the web server in step 2040. The server-side controller 5800 updates the server-side model 5600 in step 2050, and the server-side view-generating logic 5400 generates the new page in step 2060. The web server then transmits the new page to the web browser in step 2070, and the flow repeats from step 2020.

6

Now, referring back to the help desk example illustrated in accordance with FIGS. 4 and 5, and using the conventional thin-client architecture, the following steps would occur:

- (1) Consultant asks Customer for identification (ID), and enters it on the view. Browser sends ID to server.
- (2) Server validates ID and sends back HTML to refresh the screen with the addition of the customer's name and phone number.
- (3) Consultant asks customer for operating system, and enters it (e.g., NT) on the view. Browser sends screen information including operating system type to the Server.
- (4) Server sends back HTML to refresh the screen with the addition of the operating system specific table.
- (5) Consultant asks customer for operating system-specific information, such as hostname and enters it on the view. Browser sends screen information including hostname to server.
- (6) Server finds and returns the IP address corresponding to the hostname.
- (7) Consultant clicks on 'Add Another Problem' button. Browser sends screen information including add problem request to server.
- (8) Server sends screen with old problem row and additional empty problem row.
- (9) Consultant and customer have similar dialog, entering a second problem. Consultant presses Save. Browser sends screen information including save request to server.
- (10) Server saves problem reports.

Thus, entering n problem reports disadvantageously requires $3n+1$ round-trip interactions between the web browser and the web server.

When a subset of the Model is stored on the client, as in accordance with the dual-MVC approach of the present invention, a number of interactions with the server can be eliminated, thus providing better performance. For example, if the name, ID, phone, operating system, operating system-specific information, and remarks are in the client's model, i.e., are stored at the client, then: (i) the screen can be refreshed with the addition of a blank row; and (ii) the screen can be refreshed by adding the proper operating system table without going to the server.

FIG. 7 shows a generic application flow for a dual-MVC architecture according to the embodiment of the present invention as shown in FIG. 3. In step 6000, the web browser 1000 requests the initial page from the web server 5000, using the network 3000. The web server replies with the initial page in step 6010 and, in step 6020, the web browser displays the page in the user-interaction window 1200. The user interacts with the displayed page, in step 6030, by inputting data into the window. In step 6040, the client-side controller logic 1800 is executed and, in step 6050, determines whether server interaction is required. If server interaction is required, the web browser transmits the data back to the web server in step 6080. The server-side controller 5800 updates the server-side model 5600 in step 6090, and the server-side view-generating logic 5400 generates the new page in step 6100. The web server then transmits the new page to the web browser in step 6110, and the flow repeats from step 6020.

Alternatively, if the client-side logic decides that no server interaction is required, then the client-side model 1600 is updated, in step 6060, to reflect the user's input. Next, the client-side view-generating logic 1400 is executed and generates the new page in step 6070. Returning to step 6020, the

US 7,702,719 B1

7

page is displayed in the user-interaction window 1200. The application flow then repeats from step 6020.

Assuming that the name, ID, phone, operating system, operating system-specific information, and remarks are contained in the client-side model, the display of FIG. 5 can advantageously be generated without an interaction with the server. Thus, entering n problem reports requires n+2 round-trip interactions between the web browser and the web server using the dual-MVC approach of the present invention. Alternatives for improving performance often involve trade-offs. For example, if performance is critical, and the amount of space on the client's machine is large enough, more data can be copied to the client. Also, if a list of the hostnames and their corresponding IP addresses is stored on the client, the trip to the server to obtain the IP address given the hostname could also be skipped. If this were done, the number of round-trip interactions would be reduced to two. If a list of customer IDs and their corresponding names and phone numbers is stored on the client, this would eliminate the need to invoke the server to get check the customer ID and return the name and phone. All processing, except for the final save (one interaction), could then be done on the client side.

Referring to FIG. 8, a client-side frameset architecture for a dual-MVC approach according to an embodiment of the present invention is shown. Specifically, we specify an illustrative format for the client's browser in order to implement the dual-MVC approach of the invention. The client's browser screen 1000 is divided into multiple frames. There may be one or more visible frames comprising the user-interaction window 1200, an invisible library frame 1250, and an invisible application-specific frame 1300 where the model and logic to write screens resides. The library frame and the application specific frame may also be implemented in multiple frames. In any case, the library frame 1250 contains view-generating logic 1260, client-side model 1270 and controller logic 1280, each of which is independent of the application. The application-specific frame 1300 contains the view-generating logic 1400, client-side model 1600, and controller logic 1800, each of which is specific to the application.

The invention provides a library which is stored in one of the invisible frames. The library allows the developer to program the client's model, view, and controller in a manner where functions are called which are responsible for writing the HTML. This is very complicated and tedious. With the library, the developer's effort is greatly simplified.

The application frame 1300 provides a stable anchor-point for the client-side logic. Since the invention works by rewriting visible frame(s), the visible frames cannot themselves contain controller/view logic and model data, because the logic and data will be destroyed during the rewrite process.

The library frame 1250 provides an application-programming-interface (API) and implementation for generating the view in the visible frame. The API allows the programmer to specify the layout in terms of visible components, with each component having controller properties, e.g., text values, color. The implementation of the components in the library writes the appropriate information into the visible frame (e.g., HTML) to generate the described visual layout. The library frame also contains an application-independent client-side model function 1270, e.g., caching and data-movement operations. Finally, the library frame contains application-independent controller logic 1280, e.g., generic user-interface event handling.

In a preferred embodiment, the application programmer codes the application-specific frame model and controller logic using JavaScript, and generates the view via calls to the library API. The properties of the view are set from the model

8

and, when instructed to do so by the controller logic, the view renders itself into the visible frame.

Referring now to FIGS. 9A through 9C, code fragments are shown in order to illustrate a sample application using components in a library frame 1250. The application defines a simple Model (a counter value), and a number of library-frame components: (i) an LFrame, a component representing a browser window; (ii) an LForm, a component representing a user-input form in a window; (iii) an LText, a component representing a text-input field in a form; (iv) an LButton, a component representing a pushbutton on which the user may click; and (v) an LStaticText, a component representing a text label on the screen. Each of these components is implemented by code in the library frame 1250. The application-specific frame 1300 makes calls to the library frame 1250 to create components to make callbacks to the application-specific frame 1300 when certain user-input events occur, such as changes to text fields and button presses. During processing of these events, the event handlers in the application-specific frame 1300 may make calls to the components implemented by the library frame 1250 to modify their properties as a result of the user interaction. Of course, it is to be appreciated that the code sample shown in FIGS. 9A through 9C is illustrative in nature and, given the inventive teachings herein, one of ordinary skill in the art will realize various other ways to implement library frames.

Accordingly, as has been explained, the present invention provides for a dual-MVC arrangement (a client-side MVC and a server-side MVC), together with a way of implementing the two MVCs using a standard browser without applets. The invention makes the client-side MVC possible by separating the interaction window 1200 into a separate frame from the application logic (1400, 1800) and model 1600.

Further, the invention provides a client-side application-programming interface (API) for application developers to use when implementing dual-MVC applications. Using this API, which may be implemented in JavaScript, makes it easier to write applications.

Still further, the invention provides for packaging the application-independent code into a separate frame called the library frame 1250. In a preferred embodiment, the library frame contains the application-independent view-generating logic and application-independent controller logic and may contain an application-independent model. For example, it contains the logic to generate a button on the screen, and to receive events caused by clicking the button with a mouse, but does not know where to place the button, what label to place on it, or what application-specific controller action to take when the button is clicked. This is specified by the application-specific components in the application-specific frame 1300. The advantage of the library frame is that the library frame only needs to be downloaded once per browser session, because it changes only when software updates take place (e.g., over weeks or months). The browsers can thus cache the library frame for extended periods. In contrast, the interaction window 1200 is discarded/rewritten on virtually every user interaction, and the application-specific frame 1300 is reloaded on every server interaction. If the code in the library frame were moved to application-specific frame 1300, it would need to be reloaded on every server interaction, which would waste communication bandwidth and cause additional response-time delays.

Referring now to FIG. 10, a block diagram is shown illustrating a hardware implementation of a client computer system and/or server computer system suitable for implementing the dual-MVC methodologies of the present invention. As shown, both the client system and server system may be

US 7,702,719 B1

9

implemented in accordance with a processor **9000**, a memory **9010** and I/O devices **9020**. It is to be appreciated that the term “processor” as used herein is intended to include any processing device, such as, for example, one that includes a CPU (central processing unit) and/or other processing circuitry. The term “memory” as used herein is intended to include memory associated with a processor or CPU, such as, for example, RAM, ROM, a fixed memory device (e.g., hard drive), a removable memory device (e.g., diskette), flash memory, etc. In addition, the term “input/output devices” or “I/O devices” as used herein is intended to include, for example, one or more input devices, e.g., keyboard, for entering data to the processing unit, and/or one or more output devices, e.g., CRT display and/or printer, for presenting results associated with the processing unit. It is also to be understood that the term “processor” may refer to more than one processing device and that various elements associated with a processing device may be shared by other processing devices. Accordingly, software components including instructions or code for performing the methodologies of the invention, as described herein, may be stored in one or more of the associated memory devices (e.g., ROM, fixed or removable memory) and, when ready to be utilized, loaded in part or in whole (e.g., into RAM) and executed by a CPU. For example, the individual Model, View and Controller logic associated with the client and the server, as shown in FIG. 3, may be implemented in accordance with a hardware architecture as depicted in FIG. 10.

Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.

What is claimed is:

1. A method for use in a client/server system of reducing interactions between a client and server in association with an application being accessed by the client at the server, the method comprising the steps of:

configuring the server to store a model associated with the application and to execute view-generating and controller logic associated with the application; and

configuring the client to store at least a subset of the model associated with the application and to execute at least a subset of the view-generating and controller logic associated with the application;

wherein one or more portions of the application are performed at the client without the client having to interact with the server, and further wherein the client and the server both locally maintain at least a portion of the model and execute the view-generating and controller logic associated therewith.

2. The method of claim **1**, wherein the client and server communicate over a HyperText Transport Protocol network.

3. The method of claim **1**, wherein the client performs the one or more portions of the application in accordance with browser software running thereon.

4. The method of claim **3**, wherein the configuring step further comprises the step of partitioning a screen area associated with the browser software into frames.

5. The method of claim **4**, wherein the at least a subset of the model, the view-generating and the controller logic associated with the application are associated with at least one frame and one or more views for display in accordance with the application are associated with at least another frame.

10

6. The method of claim **5**, wherein the at least one view frame is a visible frame.

7. The method of claim **5**, wherein the at least one frame associated with the at least a subset of the model, the view-generating logic and the controller logic is not a visible frame.

8. The method of claim **4**, wherein the configuring step further comprises forming at least one frame with which application-independent view-generating logic and controller logic is associated.

9. The method of claim **8**, wherein the at least one application-independent view-generating logic and controller logic frame further has an application-independent model associated therewith.

10. The method of claim **8**, wherein the at least one application-independent view-generating logic and controller logic frame serves as an application programming interface for developing views to be displayed in accordance with the application.

11. The method of claim **10**, wherein the views are implemented in accordance with the HyperText Markup Language and the application programming interface is implemented in accordance with the JavaScript language.

12. The method of claim **1**, wherein the at least a subset of the model, the view-generating and the controller logic associated with the application are downloaded from the server to the client upon demand.

13. A network-based system, comprising:

a server having at least one processor operative to: (i) store a model associated with an application associated with the server; and (ii) execute view-generating and controller logic associated with the application; and

a client, coupled to the server via a network, having at least one processor operative to: (i) store at least a subset of the model associated with the application; and (ii) execute at least a subset of the view-generating and controller logic associated with the application;

wherein one or more portions of the application are performed at the client without the client having to interact with the server such that interactions between the client and server are reduced, and further wherein the client and the server both locally maintain at least a portion of the model and execute the view-generating and controller logic associated therewith.

14. The system of claim **13**, wherein the network is a HyperText Transport Protocol network.

15. The system of claim **13**, wherein the client processor performs the one or more portions of the application in accordance with browser software running thereon.

16. The system of claim **15**, wherein the client processor is further operative to partition a screen area associated with the browser software into frames.

17. The system of claim **16**, wherein the at least a subset of the model, the view-generating and the controller logic associated with the application are associated with at least one frame and one or more views for display in accordance with the application are associated with at least another frame.

18. The system of claim **17**, wherein the at least one view frame is a visible frame.

19. The system of claim **17**, wherein the at least one frame associated with the at least a subset of the model, the view-generating logic and the controller logic is not a visible frame.

20. The system of claim **16**, wherein the client processor is further operative to form at least one frame with which application-independent view-generating logic and controller logic is associated.

US 7,702,719 B1

11

21. The system of claim **20**, wherein the at least one application-independent view-generating logic and controller logic frame further has an application-independent model associated therewith.

22. The system of claim **20**, wherein the at least one application-independent view-generating logic and controller logic frame serves as an application programming interface for developing views to be displayed in accordance with the application.

23. The system of claim **22**, wherein the views are implemented in accordance with the HyperText Markup Language and the application programming interface is implemented in accordance with the JavaScript language.

24. The system of claim **13**, wherein the at least a subset of the model, the view-generating and the controller logic associated with the application are downloaded from the server to the client upon demand.

25. An article of manufacture for use in a client/server system for reducing interactions between a client and server in association with an application being accessed by the client at the server, comprising machine readable media containing one or more programs which when executed implement the steps of:

configuring the server to store a model associated with the application and to execute view-generating and controller logic associated with the application; and

configuring the client to store at least a subset of the model associated with the application and to execute at least a subset of the view-generating and controller logic associated with the application;

wherein one or more portions of the application are performed at the client without the client having to interact

10

15

20

25

30

12

with the server, and further wherein the client and the server both locally maintain at least a portion of the model and execute the view-generating and controller logic associated therewith.

26. A method for use in a client/server system of reducing interactions between a client and server in association with an application being accessed by the client at the server, the method comprising the steps of:

configuring the server to: (i) store a model associated with the application; (ii) execute view-generating logic associated with the application; and (iii) execute controller logic associated with the application; and

configuring the client to: (i) store at least a subset of the model associated with the application; (ii) execute at least a subset of the view-generating logic associated with the application; and (iii) execute at least a subset of the controller logic associated with the application;

wherein one or more portions of the application are performed at the client without the client having to interact with the server, and further wherein the client and the server both locally maintain at least a portion of the model and execute the view-generating and controller logic associated therewith;

further wherein, in accordance with such a dual model-view-controller arrangement, a model comprises application data, rules, and algorithms affecting the data, a view comprises a screen or window representation of a subset of the model that the application chooses to display, and a controller comprises the logic that processes user requests, and causes the model to be changed and the view to be refreshed.

* * * * *